Universidade de Aveiro Departamento de Biologia
Ano 2019

Bárbara Inês de Noronha Bastos

Regucalcin Evolution and Gene Function

Evolução da Regucalcina e Função do Gene

DECLARAÇÃO

Declaro que este relatório é integralmente da minha autoria, estando devidamente referenciadas as fontes e obras consultadas, bem como identificadas de modo claro as citações dessas obras. Não contém, por isso, qualquer tipo de plágio quer de textos publicados, qualquer que seja o meio dessa publicação, incluindo meios eletrónicos, quer de trabalhos académicos.

Universidade de Aveiro Departamento de Biologia
Ano 2019

Bárbara Inês de Noronha Bastos

Regucalcin Evolution and Gene Function

Evolução da Regucalcina e Função do Gene

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biologia Molecular e Celular, realizada sob a orientação científica do Doutor Jorge Vieira, IBMC e da Doutora Virgília Silva do Departamento de Biologia da Universidade de Aveiro.

Apoio financeiro da Fundação para a Ciência e Tecnologia (FCT), Programa Operacional Regional do Norte (Norte 2020), Portugal 2020, Fundo Europeu de Desenvolvimento Regional e Comissão de Coordenação e Desenvolvimento Regional do Norte

“Para ser grande, sê inteiro: nada
Teu exagera ou exclui.

Sê todo em cada coisa. Põe
quanto és

No mínimo que fazes.
Assim em cada lago a lua toda

Brilha, porque alta vive”

- Ricardo Reis

Dedico este trabalho à minha familia, amigos próximos e a todos os elementos do IBMC - I3S, que me acompanharam durante toda esta etapa, estou-lhes eternamente grata, quer por toda a sua dedicação e simpatia, quer por o apoio incondicional prestado.

o júri

Presidente: Professor Doutor António José de Brito Fonseca Mendes Calado, Professor Auxiliar, Universidade de Aveiro

Vogal – Arguente Principal: Doutor Luís Filipe Costa de Castro, Investigador Auxiliar, Ciimar – Centro Interdisciplinar de Investigação Marinha e Ambiental

Vogal – Orientador: Doutor Jorge Manuel de Sousa Basto Vieira, Investigador Principal, Instituto de Biologia Molecular e Celular

agradecimentos

Um enorme agradecimento, com um carinho especial ao meu orientador Jorge Vieira, por ser uma fonte eterna de conhecimento, de suporte e de paciência para todas as minhas lacunas, ao longo de todo o meu processo evolutivo. E outro agracedimento, também esse muito importante, à Cristina Vieira, por ter, acompanhado o meu trabalho e mostrar-se prestável sempre que assim era necessário.

palavras-chave

Regucalcina, Evolução, Drosophila melanogaster, Evolução fenotípica, Função.

resumo

A capacidade de síntese de ácido ascórbico tem sido perdida várias vezes em animais, sempre associada a mutações no gene terminal (*GULO*) da via de sintese de Vitamina C. Isto sugere que, outros genes da via de sinalização terão outras funções essenciais, para além da participação na síntese de Vitamin C. O estudo da *Regucalcina* (gene que está involvido no penúltimo passo da via de sinalização) em linhagens onde a *GULO f*oi perdida há bastante tempo podem dar pequenas pistas para o papel biológico deste gene, para além do seu conhecido involvimento na via da síntese de Vitamina C. Em humanos, alterações na expressão do gene da *Regucalcina* estão associadas a casos de obesidade, diabetes e cancro. Em *Drosophila melanogaster*, para além da *Regucalcina*, existe também um parálogo (*Dca*) que apresenta sinais de seleção positiva ao nível da proteina, é bastante sobreexpresso quando as moscas são aclimatizadas ao frio e apresenta níveis de expressão que estão correlacionados com a latitude. Para se entender os múltiplos papéis da *Regucalcina* e dos seus parálogos, neste projeto, foi usada uma combinação de análises evolutivas e funcionais. Na análise evolutiva, onde todos os genomas animais anotados foram usados, mostrou-se que a *Regucalcina* não está presente em todos os animais, que está a ser frequentemente duplicada em Protostómios (sugerindo subfuncionalização), mas não em Deuterostómios e que foi perdida várias vezes de forma independente. Uma relação evolutiva entre a presença/perda dos genes da *Regucalcina/GULO/SVCT* é também observável. Em *D. melanogaster*, através de estirpes de RNAi, foi demonstrado que os genes da *Regucalcin* e do *Dca* são essenciais. Por fim, uma análise bioinformática, revelou sítios de amino ácidos positivamente selecionados na tampa da proteína da Regucalcina, mas também no novo inferido domínio de interação.

keywords

Regucalcin, Evolution, Drosophila melanogaster, Phenotypic evolution, Function.

abstract

The ability to synthetize ascorbic acid has been lost multiple times in animals always due to mutations in the terminal gene (*GULO*) of the Vitamin C synthesis pathway. This suggests that the other genes of the pathway perform other essential functions besides participating in the synthesis of Vitamin C. The study of *Regucalcin* (gene involved in the penultimate step of the pathway) in lineages where *GULO* has been lost long time ago can give insight into the biological role of this gene besides its usual involvement in the Vitamin C pathway. In humans, changes in *Regucalcin* gene expression have been involved in obesity, diabetes and cancer. In *Drosophila melanogaster*, besides *Regucalcin*, there is also a paralog (*Dca*) that shows signs of positive selection at the protein level, that is greatly overexpressed when flies are cold-acclimated and shows expression levels that are correlated with latitude. In order to understand the multiple roles played by *Regucalcin* and its paralogs, in this project, a combination of evolutionary and functional analyses was used. The evolutionary analysis, where all annotated animal genomes were used, showed that the *Regucalcin* gene is not present in all animals, that it is often duplicated in Protostomes (suggesting subfunctionalization), but not in Deuterostomes and that it was lost multiple times independently. An evolutionary correlation between the presence/loss of the *Regucalcin/GULO/SVCT* genes in the animal kingdom is also observable. In *D. melanogaster*, using RNAi strains, it was demonstrated that *Regucalcin* and *Dca* are essential genes. And lastly, a bioinformatics analysis revealed positively selected amino acid sites at the Regucalcin protein lid, but also at the new putative interaction domain.

Index:

Table of figures:

List of tables:

List of abbreviations:

AA
Ascorbic acid 5
ADOPS
Automatic Detection Of Positively Selected Sites 10, 11, 14
AFP
Anterior Fat Body Protein Gene 6
ATP
Adenosine triphosphate 45
BLAST
Basic Local Alignment Search Tool 9, 10
bp
Base pairs 1
CDS
coding sequences 9, 10, 11
CPORT
Consensus Prediction Of interface 13
FUBAR
Fast Unconstrained Bayesian AppRoximation 15
GLDH
Glutamate dehydrogenase 36, 37
HADDOCK
High Ambiguity Driven protein-protein DOCKing 13
I3s
Instituto de Investigação e Inovação em Saúde - Universidade do Porto XII
ID
Identity 43
I-TASSER
Iterative Threading ASSEmbly Refinement 13
kD
kilodaltons 1
LRE
Luciferin-regenerating enzyme 6
NAT

Acknowledgments:

This project was born and raised in womb of the Phenotypic Evolution group team at the huge factory of ideas that is the “Instituto de Investigação e Inovação em Saúde” (i3S) as part of the conclusion of my master’s degree in Molecular and Cellular Biology at University of Aveiro. I am deeply grateful to all the lab colleagues I had the privilege to work and learn with, but especially to, Pedro Duque, for practically being a mentor, for all his patience, for dealing with me and for passing his knowledge. Also to Sara Rocha, André Sousa, Pedro Ferreira, Joel Laia and José Pimenta, for the knowledge, understanding, kindness, coaching and friendship delivered throughout the whole learning process. I am also grateful to FCUP, i3S, IBMC and UA, for giving me the opportunity to perform this work.

I would also like to save a special, massive and genuine “thank you”, to Jorge and Cristina Vieira for their knowledge, patience, the huge capability to teach and for always push my buttons a little further, for forcing me to think for myself in a scientific way, for the great guidance, and for being the humble great people that they are.

Furthermore, I also want to thank the great support of all my friends, that accompanied me during my master, so a big “thank you for your patience, time and memories” to Mariana Falcão and to Vera Martinho, it was a pleasure getting to know you and having both of you girls in my life, from that moment and also to those that have been with me since my bachelor, like Inês Oliveira. Thank you for everything, (the list would be really huge) and thank you, above all, for always having truly stayed with me, no matter what, even when we were so geographically apart.

A big thank you, to my boyfriend Diogo Sousa, for all his love and dedication, for handling all of my humor crisis during the writing process, for never giving up on me and, most importantly, for not letting me ever give up on myself. Lastly, but not least, I want to thank the support of my parents, family and close ones.

Introduction:

In the late 70's, Regucalcin (RGN) was discovered (Yamaguchi and Yamamoto, 1978). At that time it was firstly classified as a calcium (Ca^{2+})-binding protein, even though it lacked the typical EF-hand Ca^{2+}-binding motif (Shimokawa and Yamaguchi, 1993).

From a molecular biology point of view, looking at the 34 kilodaltons (kD) overall structure of the RGN protein it is possible to identify 24 β-strands forming 6 β-sheets able to bind diverse divalent cations (Ca^{2+}, Mg^{2+}, Mn^{2+} and Zn^{2+}) (Yamaguchi, 2013).

RGN as a gene, consists of six introns and seven exons (Yamaguchi, 2013) and is localized in the p11.3-q11.2 segment of the human (*Homo sapiens*) X chromosome (Marques *et al.*, 2014). Furthermore, it has an open reading frame (ORF) of 897 bp, thus coding for 299 amino acids (Misawa and Yamaguchi, 2000).

The *RGN* gene expression is regulated by numerous hormonal factors, which include calcium homeostasis pathways, calcium-regulating hormones, estrogen, insulin and other steroid hormones (represented in Figure 1). Furthermore, there are several regulatory transcription factors upstream of the 5′ flaking region that enhance the transcription of the *Regucalcin* gene, namely the AP1, NF1-A1, RGPR-p117 and β-catenin. Ca^{2+} levels can also modulate RGN expression in a process involving, calmodulin (CaM) or protein kinase C (PKC) (Marques *et al.*, 2014).

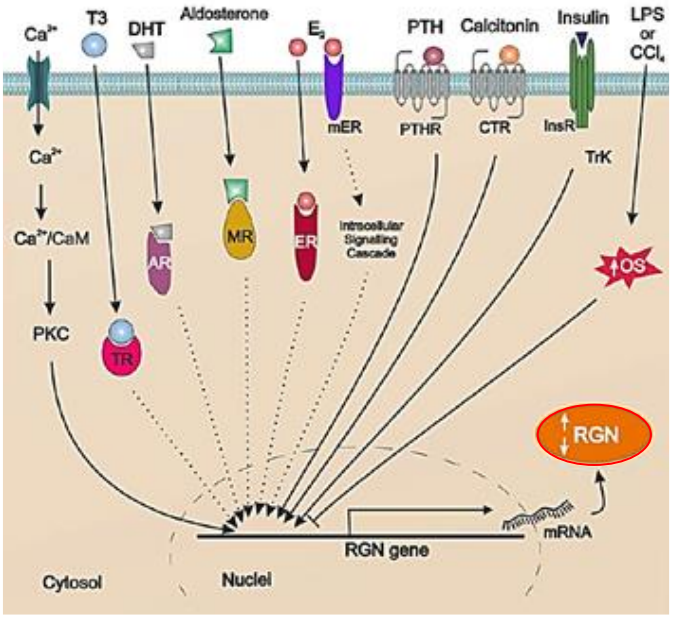


Figure 1 - Transcription factors regulating *Regucalcin* (*RGN*). Expression upregulation effects are represented with solid arrows and up/down-regulated *RGN* expression by dashed arrows. The bar-headed arrow represents inhibition. The abbreviations listed in the picture stand for: Triiodothyronine (T3), 5α-dihydrotestosterone (DHT), 17β-estradiol (E2), parathyroid hormone (PTH), lipopolysaccharide (LPS), carbon tetrachloride (CCl4), calmodulin (CaM), protein kinase C (PKC), oestrogen receptor (ER), parathyroid hormone receptor (PTHR), calcitonin receptor (CTR), insulin receptor (InsR), tyrosine kinase (TrK), thyroid hormones receptor (TR), androgen receptor (AR), mineralocorticoid receptor (MR) and oxidative stress (OS). The presented figure and description were adapted from Marques *et al.* (2014).

A study directed by (Yamaguchi and Murata, 2013) suggested that the expression of *RGN* was stimulated by insulin in liver cells both in vitro and in vivo and it is decreased in the liver of rats with type I diabetes induced by streptozotocin administration in vivo. Additionally, these authors also indicated that the overexpression of endogenous *RGN* stimulates glucose utilization and lipid production in liver cells with glucose supplementation in vitro. Furthermore, it was shown that overexpression of endogenous *Regucalcin* decreases triglyceride, total cholesterol and glycogen contents in the liver of rats, inducing hyperlipidemia (Yamaguchi and Murata, 2013). These evidences are summarized in figure 2.

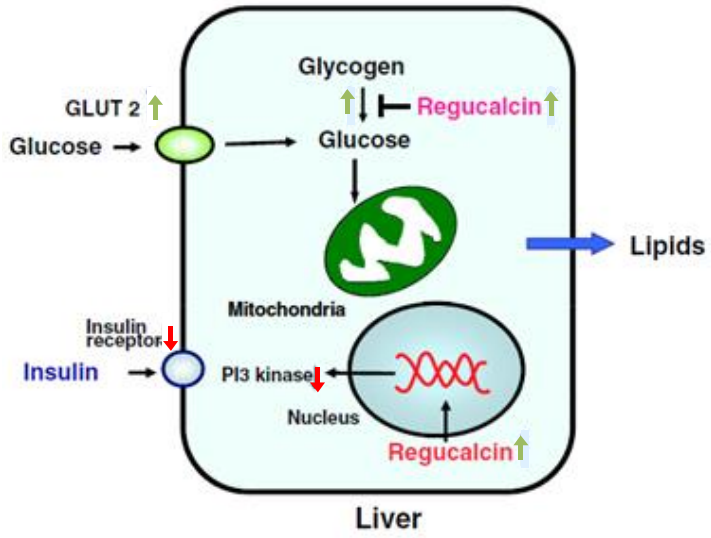


Figure 2 - *Regucalcin* overexpression mechanism inducing hyperlipidemia. This overexpression stimulates glucose utilization and lipid production in the cloned rat hepatoma H4-II-E cells. Regucalcin increases GLUT 2 mRNA expression to enhance glucose utilization in the cells, and it suppresses the gene expression of insulin receptor or PI3 kinase, which is enhanced after culture with insulin and/or glucose supplementation. Thus, overexpression of *RGN* induces insulin resistance. The presented figure and description were adapted from Yamaguchi and Murata (2013).

Moreover, nuclear *Regucalcin* also plays a role in cell proliferation and apoptotic cell death, and has also been found to cope with carcinogenesis (Yamaguchi *et al.*, 2016).

Furthermore, Ishigami *et al.*, (2001) and Maia *et al.*, (2009) showed that its overexpression seems to cause bone loss and osteoporosis and in situations of low Regucalcin levels the glucose tolerance decreases and there is an abnormal lipid accumulation in the liver.

It was also proposed that *Regucalcin* is downregulated in human prostate and breast cancers (Ishigami *et al.*, 2001; Maia *et al.*, 2009). These findings are summarized in figure 3.

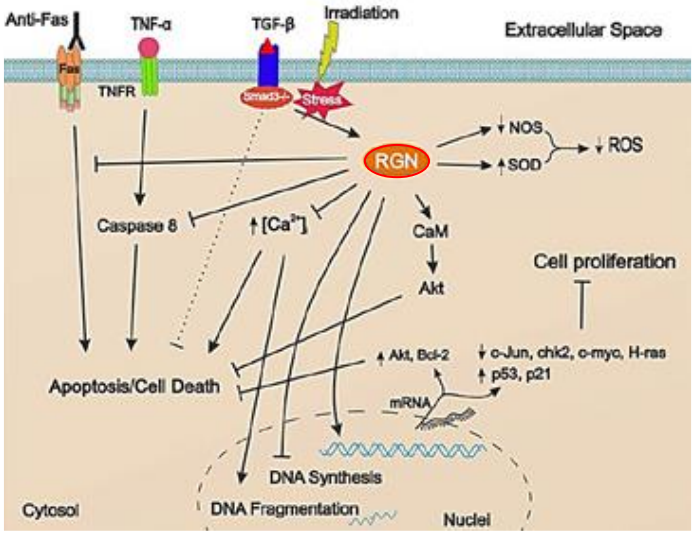


Figure 3 - RGN role in cell proliferation and apoptosis. Arrows indicate activation and bar-headed arrows represent inhibition. RGN diminishes the production of ROS, blocks the increase of intracellular calcium, inhibits caspase 8 activity, enhances activity of Akt pathway and increases the expression of apoptosis inhibitors Akt-1 and Bcl-2 leading to inhibition of apoptosis. RGN also blocks apoptosis induced by Fas system. Dashed bar-headed arrow indicates the inhibition of apoptosis in Smad 3 knock-out animals concomitant with increased levels of RGN. In turn, RGN increases the transcription of p53 and p21, while repressing the expression of *c-Jun, chk2, c-myc* and *H-ras* genes, thus blocking cell proliferation. Presented figure and description were adapted from Marques *et al.* (2014).

The RGN protein, works as a suppressor of multi-signaling pathways in several types of cells and tissues. Regucalcin has an extremely conserved amino acid sequence (70–90% identity) in vertebrates (Misawa and Yamaguchi, 2000), strongly suggestive of essential biological functions (Scott and Bahnson, 2011) and was even posteriorly named, in mammals, the senescence marker protein-30 (SMP30), which was firstly described by Amano et al., (2014) as an age-associated protein.

RGN protein can be translocated from the cytoplasm to the nucleus in various types of cells playing a role in regulating nuclear functions, like maintaining intracellular calcium homeostasis which, on the other hand, activates multiple transcription factors, such as PI3K, AP-1 and b-catenin as described by (Laurentino *et al.*, 2012). Furthermore, RGN also contributes in reducing intracellular levels of oxidative stress, achieved by modulating the activity of enzymes involved in generation of oxidative stress as well as in the antioxidant defense system (Marques *et al.*, 2014). Moreover, it inhibits some protein kinases, like protein kinase C (PKC) , protein phosphatases and protein synthesis in the cytoplasm, as well as nuclear DNA and RNA synthesis (Nakajima, Murata and Yamaguchi, 1999). The described RGN functions can be seen in figure 4.

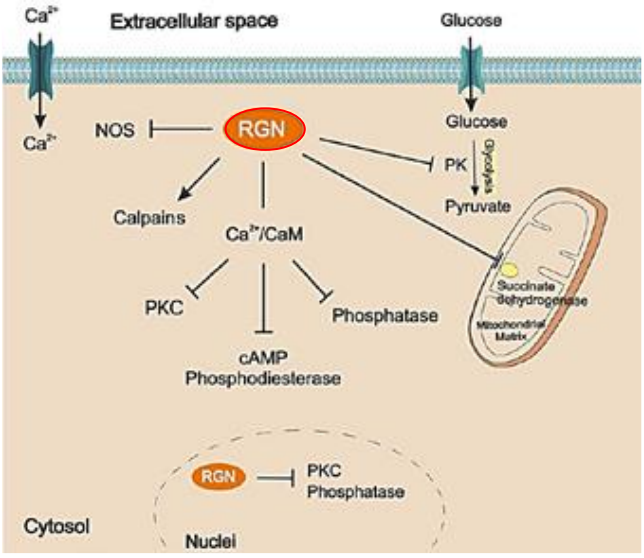


Figure 4 - RGN role in intracellular signalling and metabolism. Solid arrows indicate activation by RGN and bar-headed arrows represent inhibition. RGN decreases NOS (Nitric Oxide Synthase), PK (Protein Kinase) and succinate dehydrogenase enzymes activity. Regucalcin also inhibits Ca2+/CaM dependent activation of PKC (protein kinase C), cAMP phosphodiesterase and phosphatases. NOS, PK, CaM (calmodulin), PKC. Presented figure and description were both adapted from Marques *et al*. (2014).

Additionally, RGN is also known to be involved in ascorbic acid (AA, also known as Vitamin C) biosynthesis, assuming a gluconolactone function in the penultimate step of the animal pathway. A study conducted by (Amano *et al.*, 2014) showed that a SMP30/GNL KO mice, a deuterostomian organism, were unable to synthesize AA in vivo. Although many functions have been described in deuterostomians, within Protostomians, this information is remarkably scarce.

Nevertheless, proteins such as the anterior fat body protein (AFP) from the flesh fly (*Sarcophaga peregrina*) show high homology with the mammalian RGN (Yamaguchi, 2011). The maximum identity of deduced amino acid sequence of anterior fat body protein (8–277 amino acid regions) from flesh fly and rat Regucalcin is 33% as showed by (Yamaguchi, 2011). AFP is expressed in the anterior pair of fat body lobes of last-instar larvae and in larval hemocytes and interacts with the hexamerin receptor as explained by (Vierstraete *et al.*, 2003). Some protostomians, such as the disc abalone (*Haliotis discus*) shares a much higher amino acid identity with the vertebrate Regucalcin (42-45%) (Nikapitiya *et al.*, 2008). But most insects share in between 32% to 38% (Gomi, Hirokawa and Kajiyama, 2002), 33% identity was observed by (Nikapitiya *et al.*, 2008) between disc abalone (HdReg) and flesh fly AFP. This was an interesting discovery, since the expression of *Regucalcin* mRNA in disc abalone abductor muscle was constitutive and specifically up regulated after calcium administration (Nikapitiya *et al.*, 2008).

As indicated by Yamaguchi, (2011) most of the coding sequences of vertebrate *Regucalcin* were 299 amino acids long such as chicken, rat, mouse, and human (Yamaguchi (2011).

Nevertheless, HdReg protein contained 305 amino acid residues while *Aspergilus Fumigatus* protein contained 281 amino acids and *D. melanogaster* 303 amino acids. In *D. melanogaster*, the identified RGN is present in the hemolymph and has high similarity with the AFP of *S. peregrina*.

The expression profile of *Regucalcin* is similar to the *anterior fat body protein (AFP)* gene in *D. melanogaster*, and the shared similarity suggests that these genes are indeed orthologous. (both members of the SMP-30/Gluconolactonase/LRE-like gene family (Arboleda-Bustos and Segarra 2011).

The *Drosophila cold acclimation gene* (*Dca*) is known to be a gene duplicate of *Regucalcin* in the Sophophora subgenus, which plays a role in the adaptive response to low temperatures. This gene is known to be upregulated at the transcription level when *D. melanogaster* flies are kept 1 day at the temperature of 15ºC (Arboleda-Bustos and Segarra, 2011).

Conversely, only *Regucalcin* is present in the species of the Drosophila subgenus (*D. grimshawi, D. virilis, and D. mojavensis*). Phylogenetic analysis and the molecular organization of *Dca* indicate that this is a nested intronic gene, which appeared after a duplication event of the *Regucalcin* gene. This duplication event dates after the divergence of the *Sophophora* and *Drosophila* subgenera, yet previous to the *Sophophora* divergence (Arboleda-Bustos and Segarra, 2011). Afterwards, nonsynonymous fixation increased in the branch leading to *Dca*, but not in the one leading to *Regucalcin*, suggesting neofunctionalization of the former duplicate (Arboleda-Bustos and Segarra, 2011).

Molecular evolution of *Dca* has been affected mostly by its implication in the adaptive response to cold temperatures. Indeed, the gene has evolved under stronger purifying selection in the temperate climates, when compared with the tropical *Sophophora* species, as revealed by the ratio of nonsynonymous to synonymous substitutions. This result is consistent with functional constraints acting on the Dca protein to keep species adaptation to temperate climates (Arboleda-Bustos and Segarra, 2011). As such, as suggested by Lee *et al.*, (2011), *Regucalcin* likely kept its ancestral gene function, while *Dca* possibly acquired new functions, which can be related to Ca^{2+} homeostasis and cold acclimation.

Dca is also implicated in other biological functions, such as the regulation of wing size, in which it seems to be responsible for 5–10% of the natural wing size variation in *D. melanogaster* (McKechnie *et al.*, 2010), due to negative regulation in a sex-dependent fashion (Lee *et al.*, 2011).

Additionally, Lee *et al.*, (2011) also found a *Dca* expression cline, consistent with its proposed functional roles in size control, and also presented an insertion allele (*Dca247*) that positively influenced cell number but not cell size (using random individuals from an independent mid-cline population). Furthermore these authors also extrapolate that *Dca* might be involved in thermal tolerance, diapause, and body size control. In summary, all of these findings are suggestive

of neofunctionalization of the *Dca* gene, which is acquiring functions important for climatic adaptation.

It should be noted that the accumulation of mutations in genes resulting from duplication events, is the main source of new genes appearance. Indeed, an entirely redundant duplicate copy will not be kept in a genome for long periods, due to the existence of deleterious mutations, which accumulate over time and turn that gene nonfunctional (Zhang, 2003). A functional divergence is needed between duplicates for their long-term genome retention (He and Zhang, 2005). The neofunctionalization (NF) hypothesis of Ohno, (2014) says that after duplication, one descendant gene retains the ancestral function, whereas the other acquires new functions, experiencing a period of complete functional relaxation, behaving like a pseudogene. As an alternative to this theory, there is the subfunctionalization (SF), which implies that duplicate genes experience degenerate mutations that reduce their joint levels and patterns of activity to that of the single ancestral gene (He and Zhang, 2005).

Both of these theories can work individually, but the general consensus regards the subneofunctionalization theory (SNF), in which the gene most likely acquiring new function may retain all (NF-I), none (NF-II), or some (NF-III) of the ancestral functions (He and Zhang, 2005). A summary of the current theories can be observed in figure 5.

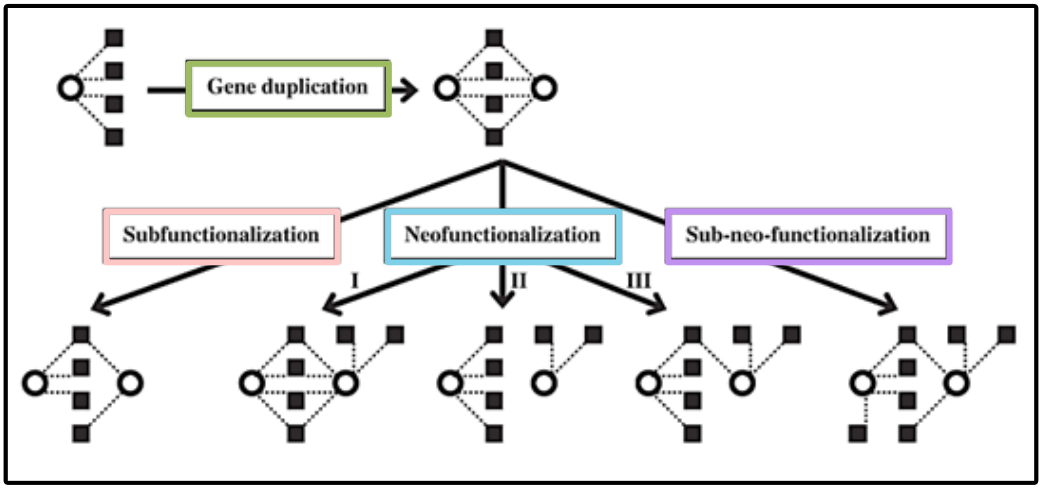


Figure 5 - Evolutionary models of functional divergence between duplicate genes. The three neofunctionalization (NF) models differ in the number of ancestral functions retained by the gene which acquires new functions. The proposed subneofunctionalization (SNF) model is a mix of NF and subfunctionalization (SF). The NF model is represented by NF-II. Duplicate genes are depicted by open circles and different gene functions are shown by solid squares. Dotted lines link genes with their functions. Presented figure and description were both adapted from He and Zhang, (2005).

Krebs, Goldstein and Kilpatrick, (2018) concluded that essential genes are conserved and ancient. Younger genes, which exist in only one or a few species, are considered less important and assume fewer organismal functions (Krylov, 2003).

The way in which essential genes emerge and how new genes accumulate essential functions is not very clear yet, since new genes arise in a continuous way through various mechanisms, such as DNA-based duplication, retroposition, and de novo origination (Kaessmann, Vinckenbosch and Long, 2009), but when they first appear, new genes should be nonessential, given that their immediate ancestral species was able to survive without them (see figure 6).

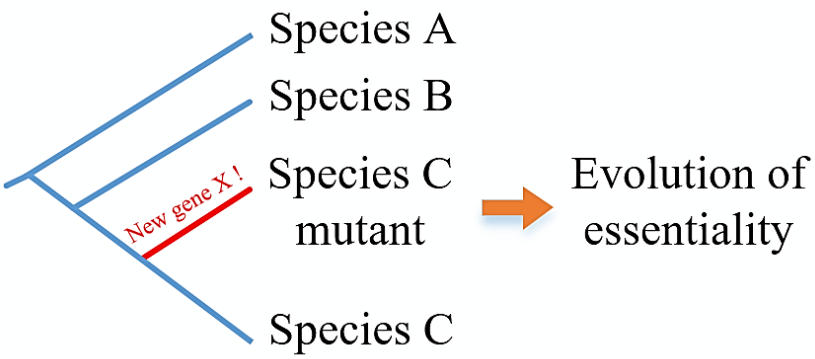


Figure 6 - Origin of new essential genes during recent evolution in Drosophila - hypothesis for the origin of a new essential gene, being the ancestral of species C immediately before the new gene X originated. Presented figure and description were both adapted from Chen, Zhang and Long, (2010).

However, little is known about their phenotypes and degrees of essentiality (Chen, Zhang and Long, 2010). A “de novo gene” has to evolve essentiality through neofunctionalization because it has no ancestral template. A duplicated gene is generated from an ancestral copy of its parental gene. And this may become essential from the loss of parents, from the switch of essentiality of paralogs, or through subfunctionalization (Lynch *et al.*, 2001). All these cases likely resulting in essential functions in a short amount of time, although through considerable different mechanisms that can correlate with very distinct degrees of essentiality.

The conclusion taken by Chen, Zhang and Long, (2010) is that the vast majority of the young essential genes have evident older and conserved paralogs and experience rapid sequence evolution. The prevalent gene structure renovation, combined with the independence between parental gene essentiality and new gene essentiality, support the neofunctionalization origin of essentiality for most new protein-coding genes, many of which may contribute to the lineage-specific developmental program.

Consequently, there are still some remaining questions left unanswered. What are in fact the functions of *Regucalcin*, in Protostomians, and are they analogous to the ones described in Deuterostomians? Did the *Dca* gene undergone a process of subfunctionalization or neofunctionalization?

The aim of this study was, therefore to fully understand the evolutionary history of the *Regucalcin* gene, while correlating putative protostomian *RGN* gene functions with the ones already described. The relevance of understanding the unknown roles of the *Regucalcin* gene within the animal kingdom could be a turnover in many more studies to come, being this gene involved with so many biochemical pathways and therefore crucial to life and health.
To do so, the use of bioinformatics was key in order to use as much as possible all available data.

Materials and Methods:

1. Animal *Regucalcin* phylogenies

The coding sequences (CDS) files were transferred from NCBI (https://www.ncbi.nlm.nih.gov/assembly/) by typing "Animals" under the "Assembly" search option. Given the incomplete overlap in the CDS annotations between the GenBank and RefSeq databases, all of the available data in FASTA format from both was downloaded, hoping to obtain the maximum information possible. Then, using the SEquence DAtaset builder (SEDA) (http://sing-group.org/seda/) software “NCBI Rename” option, a prefix was added to each file name, with information on the species name, common name, and kingdom to which the species belonged to.
This phase permitted to identify contaminations with species misclassified as animals in the downloaded files, as it was the case for *Escherichia coli* (bacteria) and for the *Bovine orthopneumovirus* (virus). The sequences representative of three species were posteriorly deleted from the dataset. Due to the cheer size of the animal complete CDS FASTA files, specifically 6.3 and 26.4 GB for GenBank and Refseq, respectively, it was necessary to narrow the information, focusing on the gene of interest, *Regucalcin*.

For this purpose, a Blastn search was performed using the SEDA software by selecting the *Homo sapiens* Regucalcin protein available at NCBI (NP_690608.1) as query against the GenBank and RefSeq CDS files, separately. The BLAST algorithm version used was 2.7.1+ and the Blastn parameters selected included a 0.05 expect value (E-value) and a limitless number of BLAST hits to

retrieve. These output files were further processed using SEDA's "NCBI rename" option, to prefix the header of each of the retrieved sequences with the name of the species, common name, and the family name to which the species belongs to. For both GenBank and RefSeq data, SEDA´s “Merge” option was used so the individual species files could be merged into a single file. The GenBank and RefSeq files were then processed for the removal of sequence line breaks using the "Reformat file" option. Using the "Rename header” option, the sequence headers were changed to keep only the species name, common name, family name, and protein accession number. Again using the "Merge” option, the processed GenBank and RefSeq files were merged into a single file.

By using the "Remove redundant sequences", option, identical nucleotide sequences were removed and the corresponding sequence headers merged. For this reason, some of the sequences represented in the final phylogeny show information for more than one species (Supplementary material, table 13).

The sequences in the processed FASTA file were then aligned using the "MUSCLE (Codons)" option present in the MEGA7 (Kumar, Stecher and Tamura, 2016) software. Still in MEGA7, and using this aligned sequence file, a neighbor-joining phylogeny was obtained using the standard parameters to help with the identification of possible CDS isoforms in need of removal. The potential isoforms detected were analysed by protein sequence comparison, using the "Align two or more sequences" option in a standard protein BLAST available at NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins). The identification of isoforms was performed following these criteria: 98% or more similarity between sequences, sequences with less than 98% similarity but with obvious annotation errors (such as wrong intron locations) or sequences already identified with an "Isoform" tag in by the NCBI database. In the cases of 100% similarity and length between isoforms, the isoform to remove was chosen in a random way. In the remaining cases, isoforms chosen for removal were the least similar to *Homo sapiens* Regucalcin between the compared lot, in terms of size and/or identity. The isoforms identified and removed in the *Regucalcin* dataset can be seen in (Supplementary material, table 17).

A *Regucalcin* Bayesian phylogenetic tree was created by analyzing the FASTA format file that resulted from the isoform removing step, using the Automatic Detection Of Positively Selected Sites (ADOPS) pipeline (Reboiro-Jato *et al.*, 2012). In this pipeline, nucleotide sequences were firstly translated and aligned using the amino-acid alignment as a guide. MUSCLE alignment algorithm was used as implemented in T-Coffee (Notredame, Higgins and Heringa, 2000). Only codons with a support value above two were used for phylogenetic reconstruction. MrBayes 3.1.2 (Ronquist *et al.*, 2012) was used as implemented in the ADOPS pipeline. The general time-reversible model (GTR) model of sequence evolution was implemented in the analysis, allowing for among-site rate variation and a proportion of invariable sites. Third codon positions were allowed to have a gamma distribution

shape parameter different from that of first and second codon positions. Two independent runs of 1,000,000 generations with four chains each (one cold and three heated chains) were performed. The average standard deviation of split frequencies was always around 0.01 and the potential scale reduction factor for every parameter about 1.00, showing that convergence has been achieved. Trees were sampled every 100th generation with a defined burn-in of 25% for the complete analysis (first 2500 samples were discarded). The non-discarded trees were used to compute the Bayesian posterior probability values of each clade of the consensus tree. The Nexus format Bayesian trees produced as output by the ADOPS pipeline were converted to Newick´s format using the Format Conversion Website (http://phylogeny.lirmm.fr/phylo_cgi/data_converter.cgi). This Newick formatted file was imported to MEGA7 in order to root the consensus phylogenetic tree.

It was observed that, this Bayesian phylogenetic tree did not present well-defined branches. So, by manually verifying the produced alignment files (using MEGA7 (Tamura *et al.*, 2007)), it was possible to see that some of the CDS sequences reduced the amount of information gathered by MrBayes for the phylogenetic relations inference due to the presence of alignment gaps. By direct observation of the aligned sequence (using MEGA7 (Tamura *et al.*, 2007)), the problematic ones were then removed manually and are displayed in (Supplementary, tables 14 to 16). By re-doing the ADOPS pipeline protocol using the refined FASTA format file, a much more defined Bayesian phylogenetic tree was obtained from which more accurate observations could be made.

Hence, a minimum of three species representative of a given taxonomic group in which the *Regucalcin* gene with all expected features was missing was used to sustain the hypothesis of *Regucalcin* gene loss in their respective lineage.

Using these criteria, a higher confidence level could be achieved regarding the obtained results, since it is very unlikely that technical issues affect three different species genomes in the same way.

2. Fly experiments

2.1. Drosophila melanogaster Strains maintenance

To analyse the expression of *Regucalcin* and of *Dca*, different fly stocks were used namely, a standard transgenic actin 5C GAL4-UAS driver (25374) from Bloomington Drosophila Stock Center (https://bdsc.indiana.edu/), and RNAi strains for *Dca* (103377) and *Regucalcin* (105509).

These fly stocks were kept at environmental chambers at a continuous temperature of 25ºC with 12h day/night cycles. Flies were reared on cornmeal food supplemented with yeast extract, without the presence of Vitamin C on its composition.

2.2. Fly crossing and sample acquisition

Crossings were performed in cornmeal food tubes with the driver (GAL4) against *Dca* and *Regucalcin* RNAi strains, in both directions (♂ driver x ♀ RNAi; ♀ driver x ♂ RNAi). Each crossing was transferred to new food tubes every two days, until a total of 6 tubes was used, being the original progenitors discarded from the last used tube after 2 days. The crossing tubes were then kept at the temperature of 25ºC.

The resulting progeny was categorized and consecutively separated and counted, by phenotype and gender, at the day of birth. Flies that presented curly wings do not express RNAi, according to the used stocks specifications, and thus were immediately discarded.

3. Protein Structure and interaction domain prediction

The interactome for the Regucalcin protein in *D. melanogaster* was obtained by using the EvoPPI web tool (http://evoppi.sing-group.org/dashboard), which allows the comparison of publicly available data from the main Protein-Protein Interaction (PPI) databases for distinct species (Vázquez *et al.*, 2019). In the "query" option, the "same species" parameter was chosen, and "*Drosophila melanogaster"* was selected. The "gene" field selected was *Regucalcin* (Gene ID: 32164).

From the resulting interactome, some proteins stood out due to their calcium levels regulation functionalities, namely, the Chloride intracellular channel (Gene ID: 32349), the Annexin B10 (Gene ID: 33019), the FK506-binding protein 14 (Gene ID: 37449), the Supercoiling factor (Gene ID: 38145), the Calreticulin (Gene ID: 41166), the Translationally controlled tumor protein (Gene ID: 41341) and the Seipin (Gene ID: 31245). Other two were also analysed due to their high expression only in female pupal developmental stages (due the gender selection in fly experiments), namely CG11267 (Gene ID: 39476) and CG2862 (Gene ID: 33471). The Glutathione S Transferase O3 (Gene ID: 38972) protein was also considered relevant, for further analyses due to its putative participation in the Ascorbic Acid synthesis pathway, according to Flybase.

Afterwards, the National Centre for Biotechnology information (NCBI) database (https://www.ncbi.nlm.nih.gov/gene) was accessed to confirm the authenticity of the Gene ID and

then, the Uniprot website (https://www.uniprot.org/) was used to convert all the Gene IDs to uniprot IDs, (Q9VY78, P22465, Q9V3V2, Q9W0H8, P29413, Q9VGS2, Q9V3X4, Q9VU35, Q8STA5 and Q9VSL2 respectively). Then, to predict the structure of Regucalcin and the uncovered interactor proteins, an Iterative Threading ASSEmbly Refinement (I-TASSER, https://zhanglab.ccmb.med.umich.edu/I-TASSER/) analysis was performed.

I-TASSER is a hierarchical approach to protein structure and function prediction, which firstly identifies structural templates from the PDB. After obtaining all the protein model structures, the ones with the highest positive C-score were chosen, since higher C-score values indicate structures with a higher confidence level, within the expected value range of [-5,2] (Zhang, 2008; Roy, Kucukural and Zhang, 2010; Yang *et al.*, 2015) .

Using the inferred protein structures, it was possible to predict active (or potential interacting) and passive residues, within the amino acid sequences. For that purpose, the Consensus Prediction Of interface Residues in Transient complexes (CPORT, https://milou.science.uu.nl/services/CPORT/) web tool was used. This tool uses an algorithm for the prediction of protein-protein interface residues, which combines five interface prediction methods into a consensus predictor. The CPORT predictions can be used as active and passive residues in High Ambiguity Driven protein-protein DOCKing (HADDOCK), using the prediction interface (de Vries and Bonvin, 2011). As such, the CPORT output pdb files were posteriorly used in the HADDOCK 2.2 webserver (https://milou.science.uu.nl/services/HADDOCK2.2/haddockserver-prediction.html) to extrapolate putative interaction sites between Regucalcin and the candidate interacting proteins. The HADDOCK is an information-driven flexible docking approach for the modelling of biomolecular complexes, to drive the docking process. HADDOCK can deal with a large class of modelling problems including protein-protein, protein-nucleic acids and protein-ligand complexes (van Zundert *et al.*, 2016). Using the Prediction interface, the Regucalcin protein was considered the first molecule in all the analyses, and as the second molecule, the interacting proteins that were previously mentioned.

The HADDOCK standard parameters were altered regarding the structure definition in which the option “I am submitting it” was chosen and for the chain structures “all” were used. Afterwards, all the clusters with negative Z-score were chosen (the clusters of macromolecular, with the lowest Z-score values, are the best predictions) and all associated PDB structures were downloaded.

Subsequently, “Proteins, Interfaces, Structures and Assemblies” (PDBePISA) web tool (https://www.ebi.ac.uk/pdbe/pisa/) was used to analyse in detail the docking results. PDBePISA analyses structural and chemical proprieties of molecular interfaces, protein assemblies and likely dissociation patterns. These are based on physical-chemical models of macromolecular interactions

and thermodynamics. PDBePISA is the only platform that shows the interactions, at the residue level (Krissinel and Henrick, 2007). Within the PDBePISA platform, the option "coordinate file" was selected and each of the PDB models were uploaded. The option "interfaces" was selected and then the option "details". Then, on the "number of residues" results, the interface residues for "Structure 1 and Structure 2" were saved, as well as the interface "Solvent-accessible area, Å". Then the structure for each protein with higher interacting number of residues was picked, and from of it, the "Interfacing residues" table was saved. The same procedure was applied to all the interacting proteins and then compared against the reference Regucalcin sequence, in order to observe where the docking hits were located. Moreover, the colours red, orange and yellow were respectively attributed, to the highest, second highest and third highest number of docking sites on the 3D structure of the Regucalcin protein. The protein structure was mapped and painted using the PyMol software, which is a user-sponsored molecular visualization system (https://pymol.org/2/, by Schrödinger).

After seeing the results of the Bayesian phylogeny, and realizing that the 35 species of the Nematoda group were missing the *Regucalcin* gene, it was important to verify if there were any calcium homeostasis related proteins in *Caenorhabditis elegans*, or in other words, if there were any orthologues for the calcium interacting protein in *D. melanogaster*. Therefore, a Blastp and reverse Blastp were performed for each relevant *D. melanogaster* protein. After identifying the orthologues, an interactome was obtained for each protein (173314, 191634, 180769, 178997, and 172944) with the same method already described for the protein encoded by the gene. Using all the obtained interactomes, it was possible to perform an interception, using the Venny web tool (http://bioinfogp.cnb.csic.es/tools/venny/) to verify if any Regucalcin orthologous protein could be inferred.

4. Positively selected amino acid sites prediction

Positively selected amino acid sites were inferred with the use of codeML (Yang, 2007) as implemented in ADOPS (Reboiro-Jato *et al.*, 2012).

A *Regucalcin* Bayesian phylogenetic tree was created by analyzing the FASTA format file that resulted from the isoform removing step, using the Automatic Detection Of Positively Selected Sites (ADOPS) pipeline (Reboiro-Jato *et al.*, 2012). In this pipeline, nucleotide sequences were firstly translated and aligned using the amino-acid alignment as a guide. MUSCLE alignment algorithm was used as implemented in T-Coffee (Notredame, Higgins and Heringa, 2000). Only codons with a support value above two were used for phylogenetic reconstruction. MrBayes 3.1.2 (Ronquist *et al.*, 2012) was used as implemented in the ADOPS pipeline. The general time-reversible model (GTR)

model of sequence evolution was implemented in the analysis, allowing for among-site rate variation and a proportion of invariable sites. Third codon positions were allowed to have a gamma distribution shape parameter different from that of first and second codon positions. Two independent runs of 1,000,000 generations with four chains each (one cold and three heated chains) were performed. The average standard deviation of split frequencies was always around 0.01 and the potential scale reduction factor for every parameter about 1.00, showing that convergence has been achieved. Trees were sampled every 100th generation with a defined burn-in of 25% for the complete analysis (first 2500 samples were discarded). The non-discarded trees were used to compute the Bayesian posterior probability values of each clade of the consensus tree. The *Regucalcin* gene sequences coming from the 458 animal species presented in the Bayesian phylogeny were sorted into the following groups, namely Apoidea_gene2, Aves_gene2, Ciprinidae_gene_C1", fish_w/o_Cyprinidae_&_Salmonidae, Culicidae_gene3, Drosophila_regucalcin, Formicoidea_gene1, Lepidoptera_gene1, Lepidoptera_gene2, Lepidoptera_gene4, Mammalia_gene1, Mammalia_gene2, Reptilia_gene1, Reptilia_gene2, Sophophora_Dca and Sophophora_regucalcin.

In order to increase our confidence in the above inferences, another prediction method was used, namely, Fast Unconstrained Bayesian AppRoximation, (FUBAR), which is also used to infer selection (Murrell *et al.*, 2013). The sequence alignment and phylogenetic tree used were those produced by ADOPS.

All the details can be observed in the B+ database (bpositive.i3s.up.pt; "The evolution of regucalcin-like genes" (BP2018000005)). However, for the case of three datasets, namely, Aves_gene1, Mammalia_gene2 and fish_w/o_Cyprinidae_&_Salmonidae, the upload to this web platform was not possible, due to the impossibility of running codeML.

The amino acid sites showing a probability higher than 90% (for both methods) were considered as positively selected.

Results:

1. Phylogenetic analyses

1.1. Regucalcin is not present in all animals

After an extensive analysis of the Bayesian Phylogenetic tree, it was possible to infer that the *Regucalcin* gene has been lost throughout its evolution history several times and in many different groups. As such, further detailed analyses were needed, both to confirm this inference and to specify which organisms or which groups were affected.

Thus, the species represented in the preliminary Bayesian Phylogenetic tree were clustered in smaller datasets, representing ten important groups, namely: Non-Bilateria, Protosmia-Lophotrochozoa, five insect groups (Hemiptera, Coleoptera, Hymenoptera, Diptera and Lepidoptera), Protostomia Non-Lophotrochozoa, Basal deuterostomes and Vertebrates. The group phylogenies (that can be seen in the supplementary material; figures S2 to S11) were filtered regarding genome contaminations and bad annotations, so just the species without any problem are shown, with an appropriate outgroup. A brief description for each group can be found on the following pages, as well as a summary cladogram for a more extensive view.

1.1.1. Non-Bilateria

After a close analysis of the Non-Bilateria phylogeny, it is possible to identify three local duplications, in *Amphimedon queenslandica*, representative of the Porifera group. For the Placozoa and Myxozoa groups, the gene is not present in the single analysed species. As such, according to our criteria, it is not possible to assume gene presence or loss in these groups. For the Anthozoa group, the gene is present in three (*Nematostella vectensis, Acropora digitifera and Orbicella faveolata*) out of a total of five species analysed, and, as for its sister group, Hydrozoa, only one species was analysed (*Hydra vulgaris*), in which the *Regucalcin* gene was absent. (See chapter 2.1 of supplementary material, for more detailed information). Therefore, the *Regucalcin* gene was present in the common ancestral this group (figure 7).

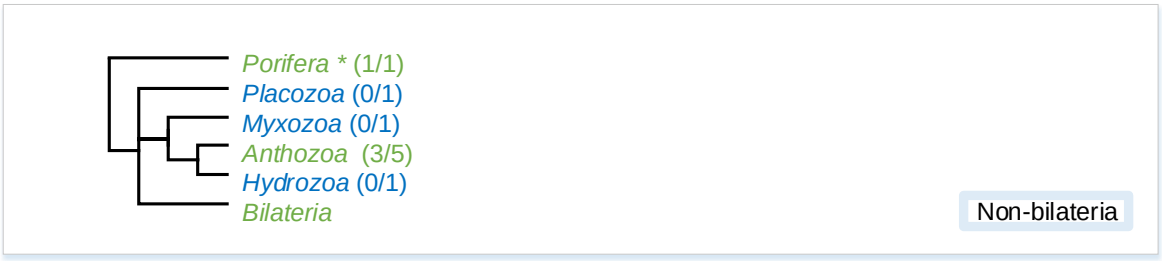


Figure 7 - Representative cladogram of *Regucalcin* gene evolution in non-bilateria species. Local gene duplications are found in one Porifera species (marked with *). Placozoa, Myxozoa and Hydrozoa species may lack a *Regucalcin* gene, but the sample size is too small (N=1) to confidently infer such gene loss. In Anthozoa it is possible to observe gene presence (painted in green). The blue colour means uncertainty about gene presence/loss. Taxonomic relationships are depicted as in Tree of life web project by Maddison, Schulz and Maddison, (2007).

1.1.2. Protostomia - Lophotrochozoa

After a close analysis of the specific Lophotrochozoa phylogeny, it is possible to verify *Regucalcin* gene presence in three species of Bivalvia (*Crassostrea gigas*, *Crassostrea virginica* and *Mizuhopecten yessoensis*) out of the four species analysed. Within the Bivalvia, in the *Crassostrea* genus lineage, it is possible to observe ancestral and local duplications, in the terminal branches, in an individual way, for both species.

As for the *Mizuhopecten yessoensis* species, local gene duplication events happened. There is also another local duplication that can be observed in this species, yet, due to the low confidence level of the tree branch (56) it is assumed that this sequence could go together with the ones previously mentioned.

Furthermore, in the Gastropoda group, three species were analysed, namely, *Aplysia californica, Lottia gigantea, Biomphalaria glabrata* and it is possible to observe local gene duplications in *Aplysia californica.*

Nevertheless, some representative sequences of Bivalvia and Gastropoda groups namely, *Crassostrea gigas* (XP_011448966.1) *and Lottia gigantea* (ESP05381.1) seem to be the result of a genome contamination. This was confirmed after a Blastp analysis, resulting on the removal of those sequences. In the Cephalopoda group, just one species was analysed, *Octopus bimaculoides*, in which local gene duplications are observable.

For the Annelida and Brachiopoda groups, two (*Capitella teleta* and *Helobdella robusta*), and one (*Lingula anatina*) species were analysed, respectively. The gene is absent in *H. robusta*, but present in *C. teleta* and *L. anatina*.

Capitella. teleta is represented by two sequences, likely derived from an ancestral duplication. *Lottia anatina* also presents two sequences that possibly derived from local duplications (See chapter 2.2 of supplementary material, for more detailed information).

Taking into account the information provided by the Lophotrochozoa group phylogeny, it was possible to infer that three gene duplications may have happened with two affecting just one lineage, giving rise to four possible genes.

The relationship established between Bivalvia, Gastropoda and Cephalopoda groups is a polytomy, as depicted from ToL by Maddison, Schulz and Maddison, (2007)) as well as the relation between Annelida and Brachiopoda. The *Regucalcin* gene was present on the common ancestral of the Lophotrochozoa. After the duplication events, some copies were independently lost, in Bivalvia genes 1, 2 and 3. The *Regucalcin* gene duplicates also may have been lost in Cephalopoda (gene 1, 2 and 3), Annelida (gene 1 and 3) and Brachiopoda (gene 1, 2 and 3). *Regucalcin* gene 1 is present in the Gastropoda, gene 2 is present in Annelida, gene 3 in Gastropoda and gene 4 in all of the analysed groups. This information can be summarized as a cladogram, based in Tree of Life web project (ToL) by Maddison, Schulz and Maddison, (2007) in figure 8.

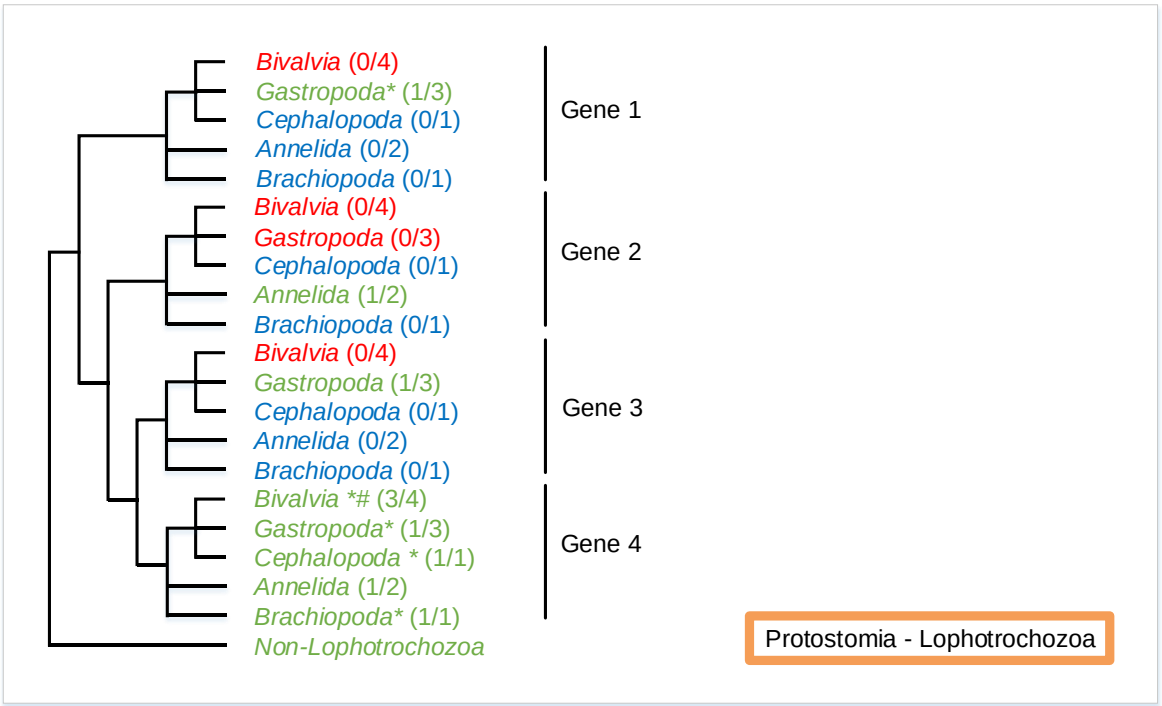


Figure 8 - Representative cladogram of *Regucalcin* gene evolution in Protostomia - Lophotrochozoa. One ancestral gene duplication happened, and one of the duplicates underwent two posterior duplication events. Local duplications that affected a single species can be identified in five groups (represented by a *) and one in one or more species can be seen (represented by a #). Gene presence is identified with the color green. Likely gene losses are highlighted in red and in blue, uncertainty regarding gene loss is represented. Taxonomic relationships are depicted as in Tree of life web project by Maddison, Schulz and Maddison, (2007).

1.1.3. Insecta – Hemiptera/Blattodea

In Hemiptera, it is possible to infer two ancestral duplication events that originated three genes, although is not possible to define the relationship between the three genes, resulting in the proposed polytomy.

The Blattodea outgroup, has only a representative species for gene 1, *Zootermopsis nevadensis*, with evidence of local duplications.

For the Pentatomidae group, the Regucalcin gene 2 is present with evidence for local duplications in *Halyomorpha halys*. Nevertheless, genes 1 and 3 are missing in this species and, as such, no inference regarding gene presence/loss can be performed in these cases for the represented group.

In its sister group, Cimicidae, it is possible to observe the presence of the gene 1 in *Cimex lectularius*. The species *Nilaparvata lugens* is the analysed representative of the Delphacidae group, in which genes 1 and 3 are present, the latter showing evidence of local duplications. For the Liviidae group, no genes can be observed in *Diaphorina citri*.

The *Regucalcin* gene was likely lost in the Aphididae group, since every copy is absent in the three analysed species, *Acyrthosiphon pisum*, *Diuraphis noxia* and *Myzus persicae*.

For the Pediculidae group, genes 1 and 2 are present in *Pediculus humanus corporis*, while gene 3 may have been lost. The phylogenetic relationships are described as by (Song, Liang and Bu, 2012; Li *et al.*, 2017) (See chapter 2.3 of supplementary material, for more detailed information). The evolution of *Regucalcin* gene in this group is summarized in figure 9.

Each of these genes appears to have a particular evolutionary history. The only remarkable loss event, common to all the genes, appears to have occurred within the Aphididae lineage, although this hypothesis could be expanded to the loss of these genes in the common ancestor of Aphididae and Liviidae. In the remaining insects, it was possible to infer a loss of genes 1 and 2.

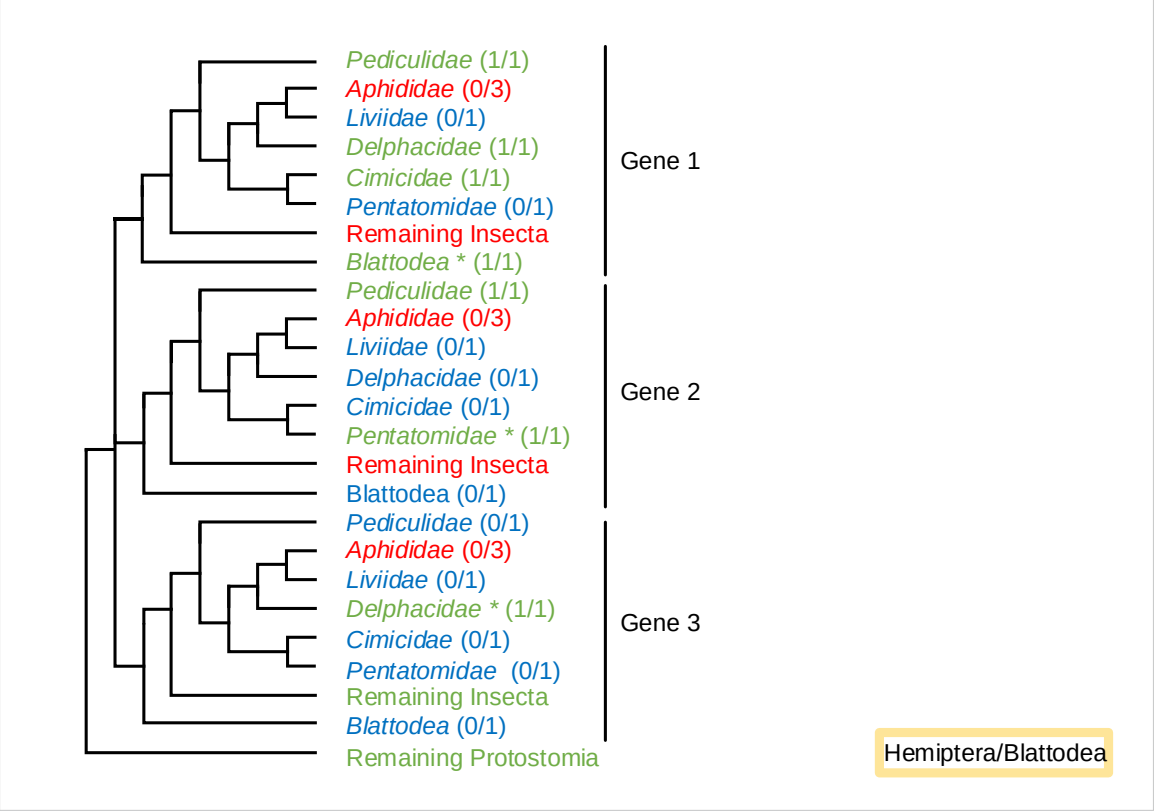


Figure 9 - Representative cladogram of the *Regucalcin* gene evolution in the Hemiptera/Blattodea. Two *Regucalcin* gene duplications are inferred to have happened at the base of the Hemiptera, giving rise to three genes. Three local duplications in a single species of a genus can also be identified (marked with a *). The green colour represents gene presence, the red means gene loss and the blue stands for uncertainty about gene presence/loss due to sample size being fewer than 3. Taxonomic relationships are depicted as in Li, H., *et al*., (2017) and Song, N., *et al.*, (2012).

1.1.4. Insecta – Coleoptera

Three gene duplications can be inferred in the Coleoptera lineage, originating four gene duplicates (genes 1 to 4). Within the Buprestidae group, only genes 1 and 3 could be found in *Agrilus planipennis*, with possible local duplication events in the latter. In the Silphidae group, it was possible to observe genes 3 and 4 in *Nicrophorus vespilloides* but not any of the four genes in *Oryctes borbonicus* from the Scarabaeidae group.

Genes 1, 2 and 4 can be found in the single representative of the Nitidulidae group, *Aethina tumida*, with local duplications of the first two genes. Regarding the Cerambycidae group, it was possible to observe genes 1, 3 and 4 in *Anoplophora glabripennis*, with notable local duplications in genes 3 and 4.

Genes 3 and 4 could be observed in the Tenebrionidae group in the single representative species, *Tribolium castaneum*, with the first presenting evidence of local duplications

Lastly, in the Curculionidae group, gene 3 could be found with likely local duplications in *Dendroctonus ponderosae,* (See chapter 2.4 of supplementary material, for more detailed information).

After gathering all the information provided by the Coleoptera group phylogeny, a specific cladogram containing more detailed information was constructed, including all the resulting information (figure 10). It is possible to infer that the *Regucalcin* gene was likely lost in the common ancestor of the groups Scarabaeidae and Silphidae independently. The same is likely to have happened to the Tenebrionidae and Curculionidae groups.

Gene 2 was likely lost on both the common ancestors of Cerambycidae and Curculionidae groups and Scarabaeidae and Silphidae. Independent gene losses are also inferred to have occurred in Tenebrionidae and Buprestidae, Nitidulidae still maintains the *Regucalcin* gene.

For gene 3, it seems that two independent gene losses have happened in Scarabaeidae and in Nitidulidae, because in the remaining groups the *Regucalcin* gene was kept.

Lastly, for gene 4, it looks like three independent gene losses might have happened in the groups: Curculionidae, Scarabaeidae and Buprestidae. As for the other four groups, the *Regucalcin* gene is still present, phylogenetic relations depicted from (Zhang *et al.*, 2018)

.

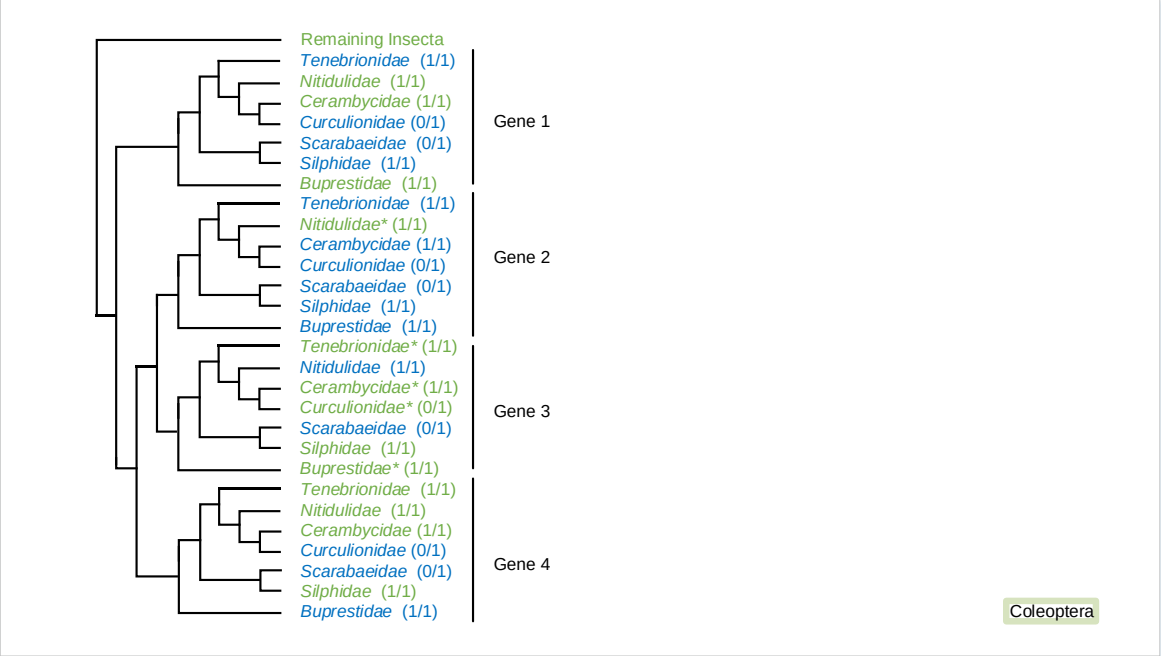


Figure 10 - Representative cladogram of *Regucalcin* gene evolution in Coleoptera. Three individual gene duplications can be inferred before the divergence of the main Coleoptera lineages. Lineages where gene copies were detected are highlighted in green, from gene 1 to gene 4. The blue color represents uncertainty regarding gene presence or loss due to small sample size (N=1) in certain taxonomic groups. Local gene duplications are marked with *. Taxonomic relationships are depicted as in Zhang *et al.*, (2018).

1.1.5. Insecta – Hymenoptera

In the Tenthredinoidea group, it is possible to observe the *Regucalcin* gene in the species *Athalia rosae* and *Neodipion lecontei*, with likely local duplications.

Within the Cephoidea group, the *Regucalcin* gene is present in *Cephus cinctus* as in the Orussidae group in *Orussus abientinus*.

A likely duplication event in the common ancestor of the Parasitoida, Vespoida, Formicoidea and Apoidea originating two duplicates, namely gene 1 and 2. Both genes are present in the Parasitoida and Apoidea groups. Gene 1 can be found in the Parasitoida in *Ceratosolen solmsi marchali* species and in Apoidea in *Ceratina calcarata* and *Eufriesea mexicana* species. As for gene 2, within the Parasitoida species, two species *Nasonia vitripennis*, *Diachasma alloeum* are represented by duplicates originated by putative local duplications, while the other two are represented by single copies.

Moreover, this gene is also present in 8 out 13 species analysed, namely *Apis cerana, Bombus terrestris, Melipona quadrifasciata, Eufriesea Mexicana, Ceratina calcarata, Habropoda laboriosa, Megachile rotundata* and *Dufourea novaeangliae*.

In the Vespoidea group, it was possible to observe gene 1 in *Polistes canadensis* e *Polistes dominula,* while a copy for gene 2 could not be traced in any of these species.

Lastly, for the Formicoidea group, gene 1 was present in 15 out of 19 species (being absent in *Dinoponera quadriceps, Linepithema humile, Monomorium pharaonis* and *Trachymyrmex zeteki*), while gene 2 was absent in all the analysed species. (See chapter 2.5 of supplementary material, for more detailed information).

After reuniting all the information provided by the Hymenoptera group phylogeny, a specific cladogram containing more detailed information was constructed, including all the resulting information (figure 11). From it, it is possible to observe that the *Regucalcin* gene is represented, overall, in all the analysed groups.

A notable gene duplication event occurred in the ancestor of the Apoidea, Formicoidea, Vespoidea and Parasitoida lineages and one of the resulting duplicates was independently lost in the Formicoidea and Vespoida.

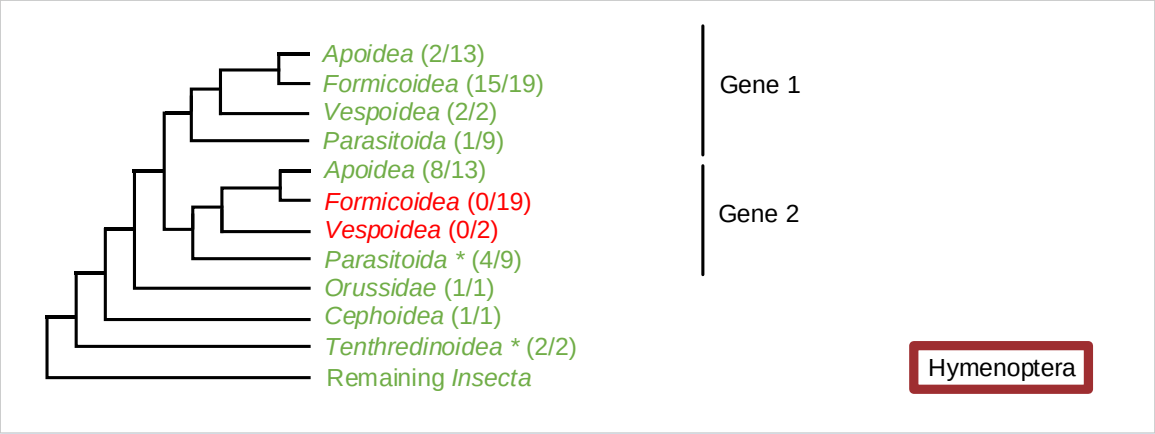


Figure 11 - Representative cladogram of the *Regucalcin* gene evolution in the Hymenoptera. A gene duplication can be inferred to have happened on the common ancestral of the Parasitoida, Vespoidea, Formicoidea and Apoidea groups. Local duplication events can be identified in Parasitoida and Tenthredinoidea (marked with a *). Two gene losses are inferred (represented in red). The green color represents gene presence. Taxonomic relationships are depicted as in Peters *et al.,* (2017).

1.1.6. Insecta – Diptera

In the Diptera order, starting with the Chironomidae group, it is possible to detect the *Regucalcin* gene in *Clunio marinus* likely affected by local duplications.

Two duplication events are likely to have happened in Culicidae group originating genes 1, 2 and 3.

Aedes aegypti, Aedes albopictus, Anopheles sinensis and *Culex quinquefasciatus* have gene 1, with evidence for local duplications in the species *A. albopictus.* The *A. albopictus* species follows this same profile in gene 2 and *A. sinensis* and *Anopheles darling* are present as a single copy. The gene 3 is present in all the six species analysed.

Two gene duplications can also be inferred for the Tephritidae group, originating genes 1, 2 and 3. The gene 1, is represented by *Bactrocera oleae* (with local duplications), *Bactrocera dorsalis* (with local duplications), *Bactrocera latifons, Rhagoletis zephyria* (with local duplications) and *Zeugodacus curcubitae.* The gene 2 consists of the species *Bactrocera oleae, Bactrocera dorsalis, Rhagoletis zephyria and Zeugodacus curcubitae.* And lastly, gene 3, is represented by the species *Bactrocera oleae, Bactrocera latifons* and *Rhagoletis zephyria*.

It was not possible to detect *Regucalcin* in the Calliphoridae group, represented by the species *Lucilia cuprina*, but it was possible to detect a *Regucalcin* sequence for one of the two Muscidae species analysed, namely *Stomoxys calcitrans.*

In the Dorsilopha group, it was possible to observe the *Regucalcin* gene in the single species, *Drosophila busckii*.

Within the Drosophila subgenus *Regucalcin* was missing in *Drosophila grimshawi*, but it is present in *Drosophila virilis, Drosophila navojoa, Drosophila mojavensis* and *Drosophila arizonae*.

In the Sophophora group, it was possible to observe the ancestral *Regucalcin* gene duplication which originated the *Dca* gene Arboleda-Bustos and Segarra, (2011). From the 21 analysed species, 17 of them had the *Regucalcin* gene and the *Dca* gene. The five species missing the *Regucalcin* gene were respectively, *Drosophila arizonae, Drosophila mojavensis, Drosophila navojoa, Drosophila erecta* and *Drosophila simulans*, while the five species missing the *Dca* gene were respectively, *Drosophila elegans, Drosophila ficusphila, D. arizonae, Drosophila kikkawai* and *D. mojavensis*. There is also an interesting fact, which is that *Drosophila ananassae* and *Drosophila bipectinata* are clustered, suffering local duplications in the latter. (See chapter 2.6 of supplementary material, for more detailed information).

After reuniting all the information provided by the Diptera group phylogeny, a specific cladogram containing more detailed information was constructed, including all the resulting

information (figure 12). It is remarkable to observe that the *Regucalcin* gene was independently duplicated many times, within the Diptera order, without the loss of the duplicate, suggesting that this duplicate may have gained important biological functions.

Figure 12 - Representative cladogram of the *Regucalcin* gene evolution in the Diptera. An ancestral duplication inside the Sophophora subgenus gave rise to the *Dca* gene, a duplicate of *Regucalcin* gene. Within lineages, gene duplications affecting a single species of genus are marked with a *, while two or more from the same genus are marked with a # (*D. ananassae* and *D. bipectinata*) Gene presence is identified with the color green. For the lineages painted in blue, there is not enough data (N<3) to surely confirm a gene loss. Taxonomic relationships are depicted as in Wiegmann *et al.* (2011).

1.1.7. Insecta - Lepidoptera

In the Lepidoptera, it is possible to infer that three duplication events were likely to have occurred, originating genes 1, 2, 3 and 4.

Regarding the Nymphalidae group, only gene 4 is represented, by *Danaus plexippus plexippus*.

Within the Pieridae group, genes 1, 2 and 4 in *Pieris rapae*. The Papilionidae group, has species representatives for all genes.

The species *Papilio polytes* and *Papilio xuthus* presents local duplications in gene 3. In gene 4 it is possible to observe grouped local duplications, for the species *Papilio polytes* and *Papilio xuthus.*

Concerning the Plutellidae group, genes 1, 2 and 4 can be observed in *Plutella xylostella*.

The Pyralidae and Bombycidae group, follow the same gene presence pattern since both of their representative species, *Amyelois transitella*, and *Bombyx mori* present the genes 2 and 4.

In the Geometridae group, only gene 4 is present in *Operophtera brumata*.

Lastly, the both analysed species of the Noctuidae group, *Helicoverpa armigera*, *Heliothis virescens* have genes 2 and 4. (See chapter 2.7 of supplementary material, for more detailed information).

After taking into account all the information provided by the Lepidoptera group phylogeny, a cladogram containing more detailed information was constructed, including all the resulting information (figure 13). Although no gene loss events could be inferred, the lack of gene 3 gene copies across several groups is noticeable. Gene 1 follows the same suggestive loss pattern, although not so drastic. As for the case of genes 2 and 4, these are highly represented in the analysed Lepidoptera groups.

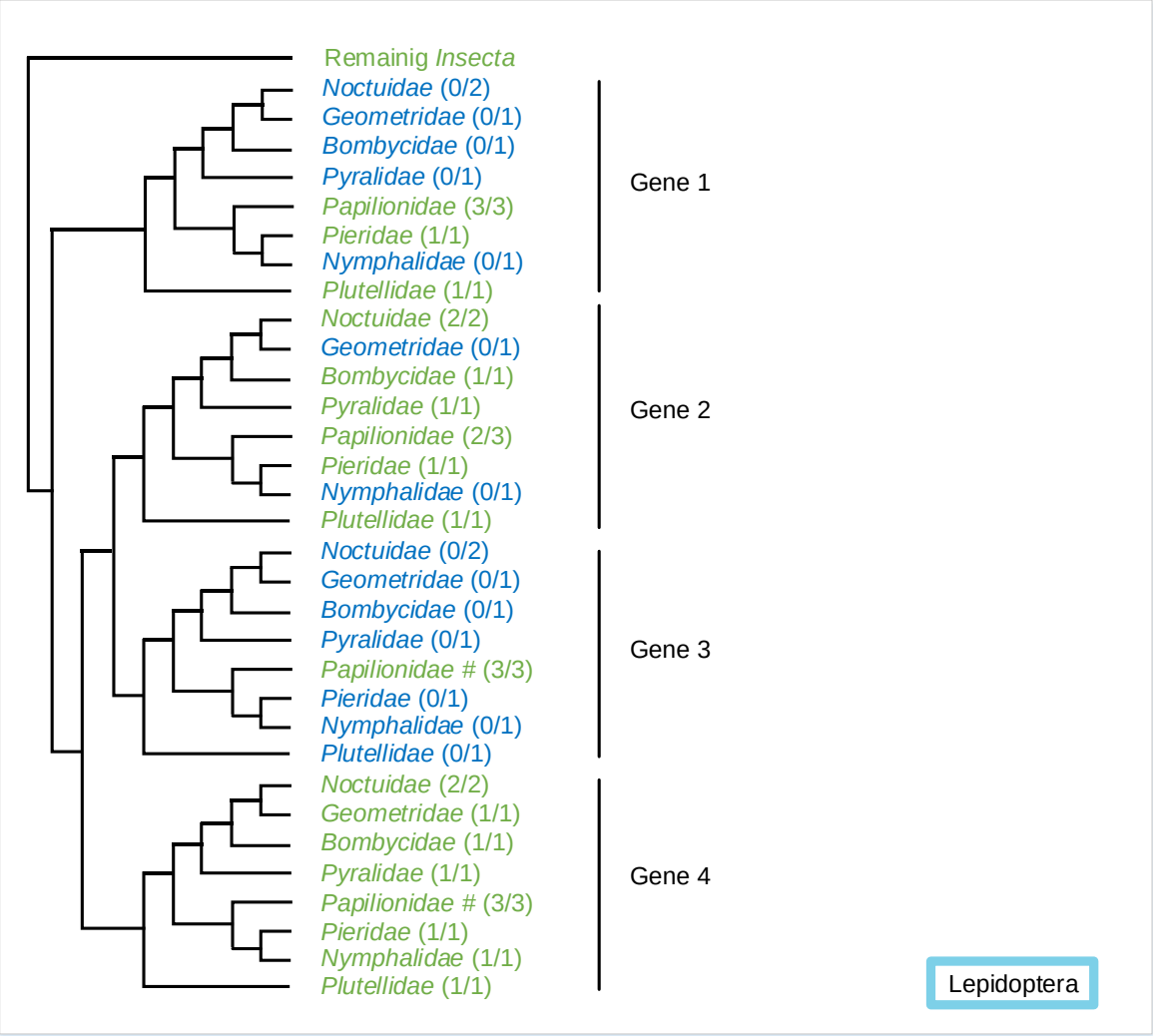


Figure 13 - Representative cladogram of the *Regucalcin* gene evolution in the Lepidoptera. Three ancestral duplication events can be inferred to have occurred, before the diversification of the represented Lepidoptera lineages. Local duplications can also be identified, in two or more species of the same genus (marked with a #) and in one species of the same genus (marked with a *). The blue colour represents lineages where gene loss could not be inferred. The green colour represents gene presence. Taxonomic relationships are depicted as in Song, F., *et al*. (2016).

1.1.8. Protostomia Non-Lophotrochozoa

Concerning the Non-Lophotrochozoa, two species from the Collembola group were analysed, namely *Folsomia candida* and *Orchesella cincta*. The *Regucalcin* gene is present and represented in *F. candida* as several copies, likely due to local duplication events. Nevertheless, one of the representative sequence of the species *Folsomia candida* (OXA60774.1) was in the wrong place in the phylogeny, and thus was removed, after showing by Blastp, that it was a genome contamination. For the case of the Crustacea group, the four studied species (*Daphnia magna, Daphina. Pulex, Armadillidium vulgare* and *Hyalella azteca*) were missing the *Regucalcin* gene. Although a representative sequence of *Hyalella azteca* species (XP_018028430.1) was present in the Bayesian phylogeny, it was likely a product of some sort of genomic contamination, shown after a Blastp analysis confirmation and being posteriorly removed.

For the Chelicerata Subphylum, without the Acari and Araneae groups, the only species analysed, (*Limulus polyphemus*) had the *Regucalcin* gene.

For the Acari group, eight species were analysed *(Ixodes scapularis, Galendromus occidentalis, Varroa destructor, Varroa jacobsoni, Tetranychus urticae, Euroglyphus maynei, Sarcoptes scabiei* and *Tropilaelaps mercedesae),* in which only three lacked the *Regucalcin* gene, namely, *Euroglyphus maynei, Sarcoptes scabiei* and *Tropilaelaps mercedesae*. It is possible to observe local duplications for *Ixodes scapularis,* and an ancestral duplication for *Tetranychus urticae.*

For the Araneae group, two species were analysed, specifically *Parasteatoda tepidariorum* and *Stegodyphus mimosarum,* and both contain the *Regucalcin* gene. Additionally, *P.tepidariorum* presents two copies of this gene, possibly originated by a local duplication event.

For the cases of Tardigrada, the two analysed species were (*Hypsibius dujardini* and *Ramazzottius varieornatus*), in Priapulida the only representative was *Priapulus caudatus*, and in Nematoda group, all the 35 analysed species have suffered a *Regucalcin* gene loss event.

As for the Platyhelminthes group, nine species were analysed and it was only possible to detect the *Regucalcin* gene in *Macrostomum lignano,* with local duplications.

Even though the fact of the sequence of the species *Opisthorchis viverrini* (OON18004.1) being present in the Bayesian phylogeny it was removed due to a likely genome contamination, after Blastp confirmation. (See chapter 2.8 of supplementary material, for more detailed information).

After reuniting all the information provided by the Non-Lophotrochozoa group phylogeny. In the Mesozoa, Priapulida and Tardigrada groups, an independent gene loss is likely to have happened.

A gene loss event was also detected in Nematoda and Crustacea groups. A Non-Lophotrochozoa specific cladogram, comprising all the result information can be observed in figure 14.

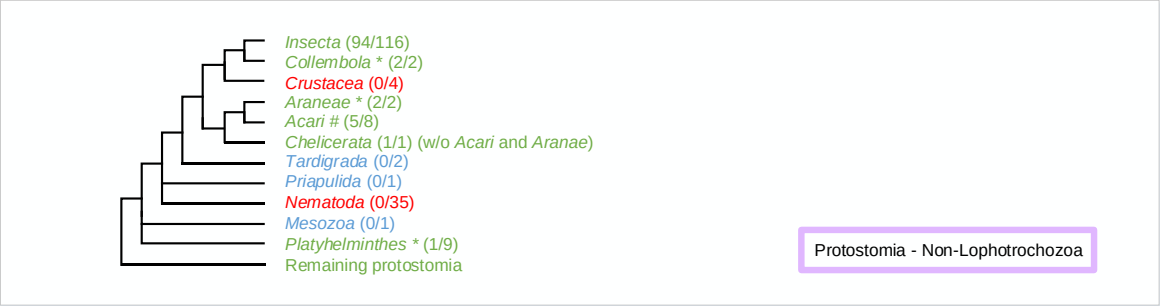


Figure 14 - Representative cladogram of *Regucalcin* gene evolution in Non-Lophotrochozoa protostomians. Four local gene duplications, likely affecting individual species (marked with a *), but also, in multiple species of the same lineage (marked with a #). Lineages where the gene was lost are painted in red. Gene presence is identified with the color green. A minimum of two independent losses are inferred. For the lineages painted in blue, there is not enough data (N<3) to surely infer a gene loss. Taxonomic relationships are depicted as in Tree of life web project by Maddison, Schulz and Maddison, (2007).

1.1.9. Basal Deuterostomians

In the basal deuterostomians, for the Acanthasteridae group, the single species analysed (*Acanthaster planci*) has the *Regucalcin* gene with noticeable local duplications.

The same scenario can be observed in the Strongylocentrotidae group for *Strongylocentrotus purpuratus.* Nevertheless, gene loss for Stichopodidae group, cannot be inferred, since a single species, *Apostichopus japonicas* was analysed.

For the Hemichordata, the *Regucalcin* gene was present in the single species analysed (*Saccoglossus kowalevskii*), affected by local duplications. Within the Urochordata group the *Regucalcin* gene was also found in *Ciona intestinalis*.

Finally, in the Cephalochordata group, *Branchiostoma belcheri* and *Branchiostoma floridae* were the identified species containing the *Regucalcin* gene. Local duplications can be observed for *B. belcheri,* while ancestral duplications that likely occurred before the split of the two species can be observed in this genus (See chapter 2.9 of supplementary material, for more detailed information).

After gathering all the information provided by the Basal Deuterostomes group phylogeny, a Basal deuterostome cladogram containing more detailed information was constructed, comprising

all the result information (figure 15). Clearly, the *Regucalcin* gene was present in the common ancestor of the basal deuterostomians.

Figure 15 - Representative cladogram of the *Regucalcin* gene evolution in Basal Deuterostomians. In the Stichopodidae lineage, *Regucalcin* loss cannot be inferred due to the sample size (N=1) (painted in blue). Within lineages, gene duplications affecting a single (marked with a *) or more than one species of a given genus (marked with a #) are also inferred in Acanthasteridae, Strongylocentrotidae, Hemichordata and Cephalochordata. The green color represents gene presence. Taxonomic relationships are depicted as in the Tree of life web project by Maddison, Schulz and Maddison, (2007).

1.1.10. Vertebrates

According, to Dehal and Boore, (2005) two rounds of whole genome duplication (WGD) events occurred at the base of vertebrates and this should be taken into account when analysing the results for *Regucalcin*.

Within the Chondrichthyes group, in which sharks and stingrays are included, the *Regucalcin* gene is present in two out of the three species analysed (*Callorhinchus milli* and *Rhincodon typus*), only being absent in *Leucoraja erinacea*.

In the Actinopteri group (without Teleostei), the *Regucalcin* gene is present in the single analysed species, *Lepisosteus oculatus*.

Within the Teleostei, it was possible to identify in the Osteoglossidae group a *Regucalcin* gene in *Scleropages formosus*.

Moreover, a duplication event in Salmonidae-like fishes and Cyprinidae-like species can be observed. In the Euteleostomorpha group the *Regucalcin* gene could be observed in 24 out of 34 species analysed.

A WGD event has likely happened in the Salmonidae group (Glasauer and Neuhauss, 2014). Nevertheless, only one *Regucalcin* gene is present in the three species analysed (gene S'), since the other (gene S") was lost.

In the Otomorpha group, four species were identified containing the gene. A WGD event also happened in the Cyprinidae group (Glasauer and Neuhauss, 2014), so it was possible to infer that the

Regucalcin gene is present in four out of the five analysed species (gene C1'), and in the five identified species (gene C1").

It is also possible to observe the *Regucalcin* gene in *Latimeria Chalumanae*, (coelacanth), a fish species that is more closely related to tetrapods (four-limbed vertebrates) than to teleost fish.

Additionally, in the Amphibia group, four species (*Nanorana parkeri*, *Xenopus laevis*, *Xenopus tropicalis* and *Rana catesbeiana*) have the *Regucalcin* gene.

Moreover, an ancestral duplication may have affected the Reptilia, Aves and Mammalia groups. Indeed two genes can be clearly identified, in these lineages (denominated as 1 and 2).

Both genes are represented in Reptilia, Aves and Mammalia with the exception of gene 1 in Mammalia. Indeed, only six mammalian species *Galeopterus variegatus* (Cynocephalidae), *Mesocricetus auratus* (Cricetidae), *Monodelphis domestica* (Didelphidae), *Phascolarctos cinereus* (Phascolarctidae), *Sarcophilus harrisii* (Dasyuridae) and *Sorex araneus* (Soricidae), have gene 1. In contrast, gene 2 is present in 100 species out of the 112 Mammalia species analysed. (See chapter 2.10 of supplementary material, for more detailed information).

After reuniting all the information provided by the Vertebrates group phylogeny, a specific cladogram is presented comprising all the results information (figure 16). On it, it is possible to observe that, two rounds of WGD have happened at the base of all vertebrates.

Given the presence of this gene in the basal deuterostomians, it was expected that the more ancestral vertebrates would have the *Regucalcin* gene. This hypothesis was confirmed due to the presence of *Regucalcin* in the Chondrichthyes.

Whole Genome Duplication events happened at the base of teleosts, although only one duplicate has likely remained functional, other notable gene loss event seem to have happened in one of the Salmonidae lineages that resulted from their specific WGD event.

Apart from these particular cases, the *Regucalcin* gene is well represented in the remaining analysed lineages.

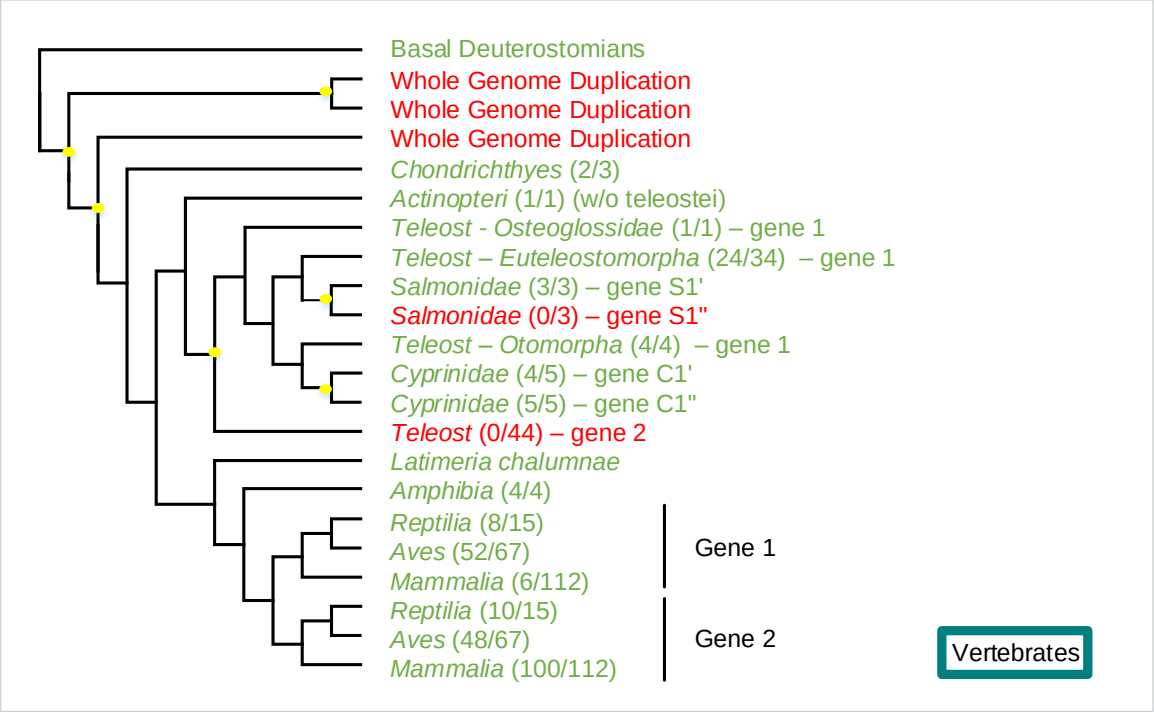


Figure 16 - Representative cladogram of the *Regucalcin* gene evolution in Vertebrates. In Vertebrates, two whole genome duplication events happened, but the duplicated genes were lost. There is also a Teleosts, Cyprinidae and Salmonidae specific Whole Genome Duplications. Moreover, an ancestral duplication between Salmonidae-like fishes and Cyprinidae-like species can be observed. These Whole genome duplications are marked with yellow dots. One out of two genes was lost (painted in red) in the Teleost and Salmonidae lineages. The *Regucalcin* gene was duplicated in the common ancestor of Reptilia, Aves and Mammalia, although gene 1 has been almost completely lost in Mammalia.The green colour represents gene presence. Taxonomic relationships are depicted as in the Tree of life web project and in Dehal. P. and Boore. J., (2005), and Glasaeur. S. and Neuhauss. S., (2014).

1.1.11. General cladograms

To have a full view of the *Regucalcin* evolution, a cladogram comprising all the animal kingdom was performed (figure 17).

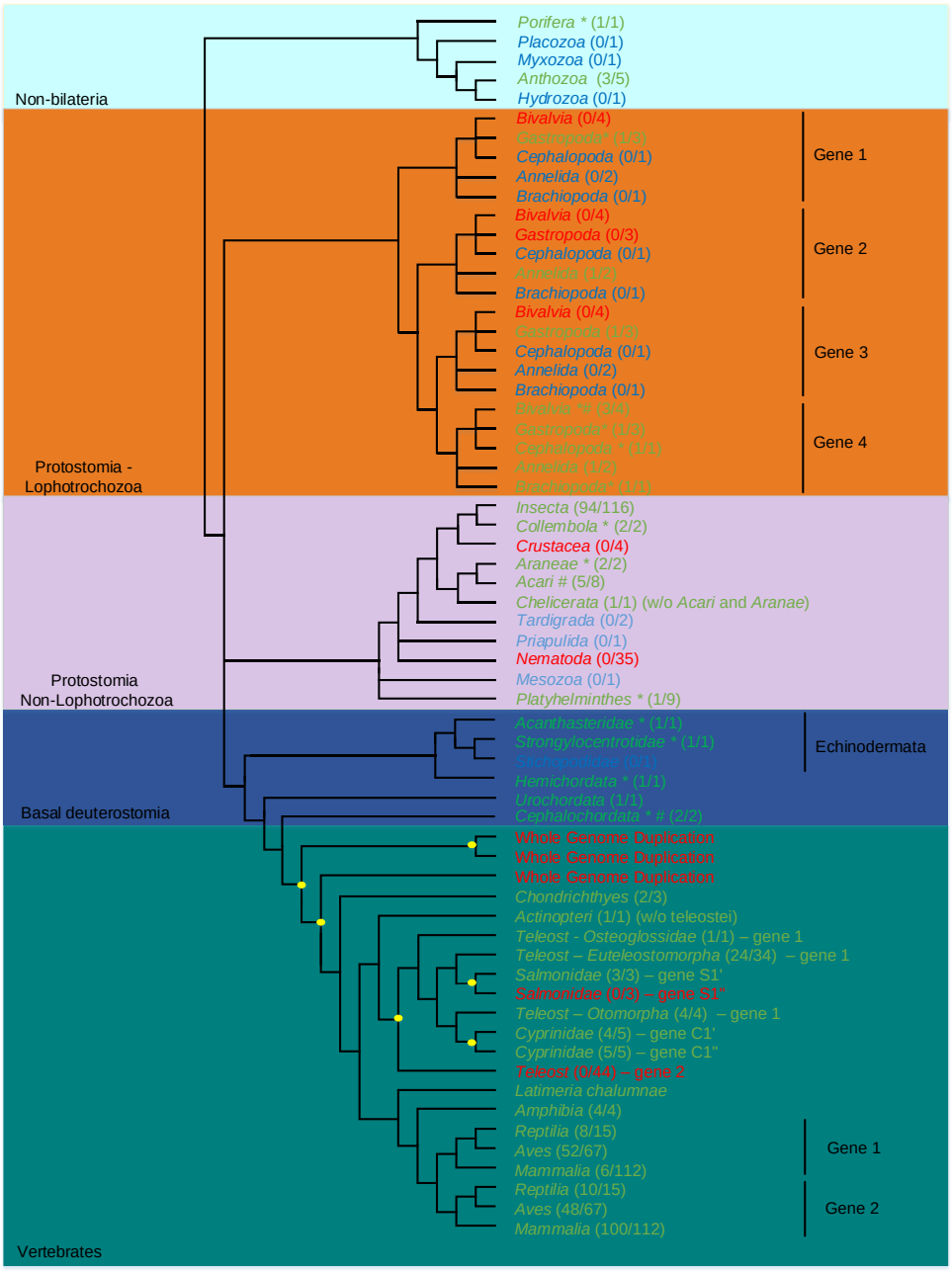


Figure 17 - General cladogram summarizing the evolutionary history of the *Regucalcin* gene. The colour green represents gene presence, the red, gene absence and the blue uncertainty, regarding gene loss/presence, due to insufficient sample size. The yellow dots represent known Whole Genome Duplication events. The (*) stands for local duplications in one species and the (#) of, at least two species of the same genus.

Neverheless, for Insecta, there is so much information available that it must be presented in a separate cladogram (figure 18). In the next sections, general *Regucalcin* evolutionary patterns are described.

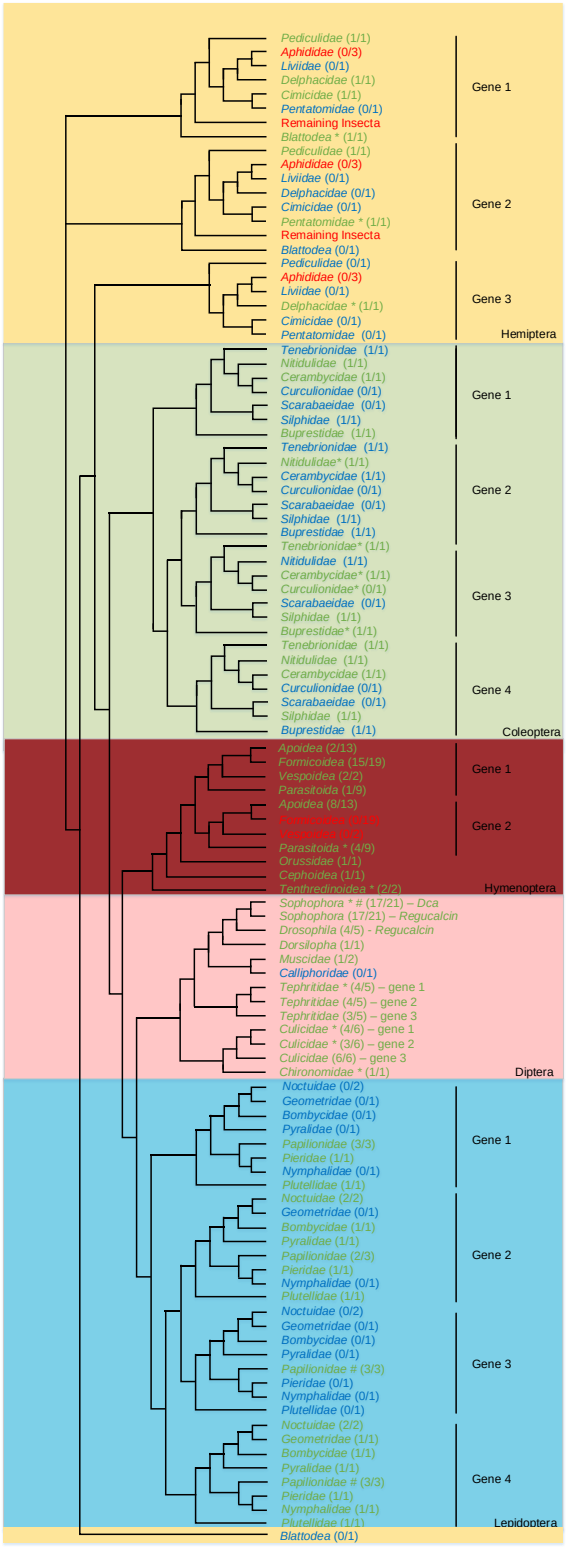


Figure 18 - General cladogram summarizing evolutionary history of *Regucalcin* gene in Insecta. The colour green represents gene presence, the red gene absence and blue for probably gene absence (due to sample size being less than three). The (*) stands for local duplications of one species and the (#) of, at least two species of the same genus. Taxonomic relationships are depicted as in Misof *et al.* (2014).

1.2. *Regucalcin* is often duplicated in Protostomes, but not in Deuterostomes

After analysing all the information in the general cladograms (figures 17 and 18), it is possible to observe that the *Regucalcin* gene is duplicated frequently in protostomes.

This is evident, for example, in Lophotrochozoans, where four duplicates originated from ancient duplications.

The same thing is happening in insects, namely in Hemiptera order, where there was an ancestral duplication event that originated three duplicates. In Coleoptera order, the *Regucalcin* gene duplicated three times, giving rise to four duplicates. A duplication event also happened inside the Hymenoptera, on the common ancestor of the groups, Parasitoidea, Vespoidea, Formicoidea and Apoidea, giving rise to two duplicates. Moreover, in the Diptera order, in the Sophophora subgenus, giving rise to *Regucalcin* and *Dca* duplicates. Finally, in the Lepidoptera order, where the *Regucalcin* gene duplicated, creating four duplicates.

Within Deuterostomes, this phenomenon does not seem to be happening so often, since the only relevant duplications observed concern the common ancestor of Mammalia, Aves and Reptilia and the case of the ancestor of the Cyprinidae. Within Protostomes, this evidence suggests possible events of subfunctionalization of the *Regucalcin* gene, where duplicate genes experience degenerate mutations that reduce their joint levels and patterns of activity to that of the single ancestral gene (He and Zhang, 2005).

1.3. *Regucalcin* gene has been lost multiple times independently

The extensive analysis summarized in the general cladograms (figures 17 and 18) shows evidence of multiple independent gene loss events. Concentrating just on figure 17, up to four gene loss events can be observed in the Lophotrochozoans, since the Bivalvia gene 1 and 3 were likely lost independently, it is possible that the observed loss for Bivalvia and Gastropoda, in gene 2, may have occurred in the common ancestor of both groups.

Then, in the Non-Lophotrochozoans, the *Regucalcin* gene was notably lost in all the analysed Nematoda and Crustacea. In Hemiptera, ancestral gene duplications originated three duplicates and all of them were lost in Pentatomindae.

In Coleoptera order, for gene 1, there might be an independent *Regucalcin* gene loss in the common ancestor of Tenebrionidae and Curculionidae, as well as in the common ancestor of Scarabaeidae and Silphicidae groups. In gene 2, an independent gene loss event happened in the common ancestor of Curculionidae and Cerambycidae groups, and Scarabaeidae and Silphidae groups and independent losses for Tenebrionidae and Buprestidae. As for gene 3, independent losses of the *Regucalcin* gene happen in Scarabaeidae and Nitidulidae. Lastly, in gene 4, Curculionidae, Scarabaeidae and Buprestidae independently lost the *Regucalcin* gene.

As for the Hymenoptera order, it is possible to infer two independent gene lost events, namely within the lineage of gene 2, in Formicoidea and Vespoidea groups.

In Diptera order, just one independent *Regucalcin* gene duplication event is inferred to have happened, namely in Calliphoridae.

Lastly of insect, is the Lepidoptera group, in which several *Regucalcin* gene independent loss events can be reported, respectively in gene 1, in the common ancestor of the groups Noctuidae, Geometridae, Bombycidae and Pyralidae, and also an independent loss event in Nymphalidae. For gene 2, there were likely independent losses in the groups Nymphalidae and Geometricidae. In gene 3, in the common ancestor of the groups Noctuidae, Geometridae, Bombycidae, Pyralidae the gene was likely lost and the same happened for Pieridae and Nymphalidae, as well as for Plutellidae.

Within the Deuterostomes, the relevant loss events appear to affect the Teleost fish (gene 2) and Salmonidae (gene S1”).

Additionally, it seems likely that all the *Regucalcin* duplicates that resulted from the two known rounds of WGD in vertebrates were lost.

Now, focusing in figure 18, in which the Insecta class can be observed with more detail. In the Hemiptera order, independent *Regucalcin* gene losses are identifiable in the common ancestor of the groups Aphididae and Liviidae both in gene 1, 2 and 3. In Delphacidae the *Regucalcin* gene is likely lost independently. In Cimicidae losses are observable for genes 2 and 3. Also, in the Pentatomidae group it is possible to identify independent losses for genes 1 and 3. Blattodea also independently lost genes 2 and 3.

1.4. Ascorbic Acid (AA) evolutionary route *(Regucalcin/GULO/SVCT)*

It is known that Regucalcin plays a role in the ascorbic acid (AA) biosynthesis, assuming a gluconolactone function in the penultimate step of the pathway (Amano *et al.*, 2014).

Commonly known as Vitamin C ($C_6H_8O_6$), the AA is a crucial co-factor for several enzymatic reactions and an indispensable antioxidant agent, stimulating the normal function and development of eukaryotic cells. Many deuterostomians are capable of synthesizing AA, while some species such as *Haplorrhini* primates, teleost fish and *Cavia porcellus* (Guinea pig) lost this ability due to the loss of the L-gulonolactone oxidase (*GULO*) gene (Drouin, Godin and Page, 2011).

It is consensual that *GULO* has been lost in the Insecta taxonomic group. However, ascorbic acid levels can be detected in one representative species from this group, the model organism *Drosophila melanogaster* (fruit fly), even when in the absence of a dietary source of this Vitamin.

Henriques *et al.*, (2019) shown that it wasn´t the fly's microbiome that was responsible for the source of ascorbic acid and that this species must synthetize AA using an alternative pathway.

The three known major pathways of ascorbic acid biosynthesis make use of distinct routes and initial substrates to synthetize an aldonolactone precursor (L-gulono-lactone or L-galactono-lactone). This is converted to ascorbic acid by either Glutamate dehydrogenase (GLDH, in the pathway used by plants and photosynthetic protists (Shigeoka, Nakano and Kitaoka, 1979; Wheeler, Jones and Smirnoff, 1998; Loewus, 1999; Wheeler *et al.*, 2015; Smirnoff, 2018) or *GULO*, in the case animal pathway (oberve figure 19). There is no *GLDH* gene in animals. *SVCT*s, are responsible for AA transport. These are in fact surface glycoproteins that belong to the nucleobase-ascorbate transporter (NAT) protein family, which can be divided into three distinct groups, given their corresponding substrate specificity: a) xanthine and uric acid, b) uracil or c) ascorbic acid (Bürzle *et al.*, 2013).

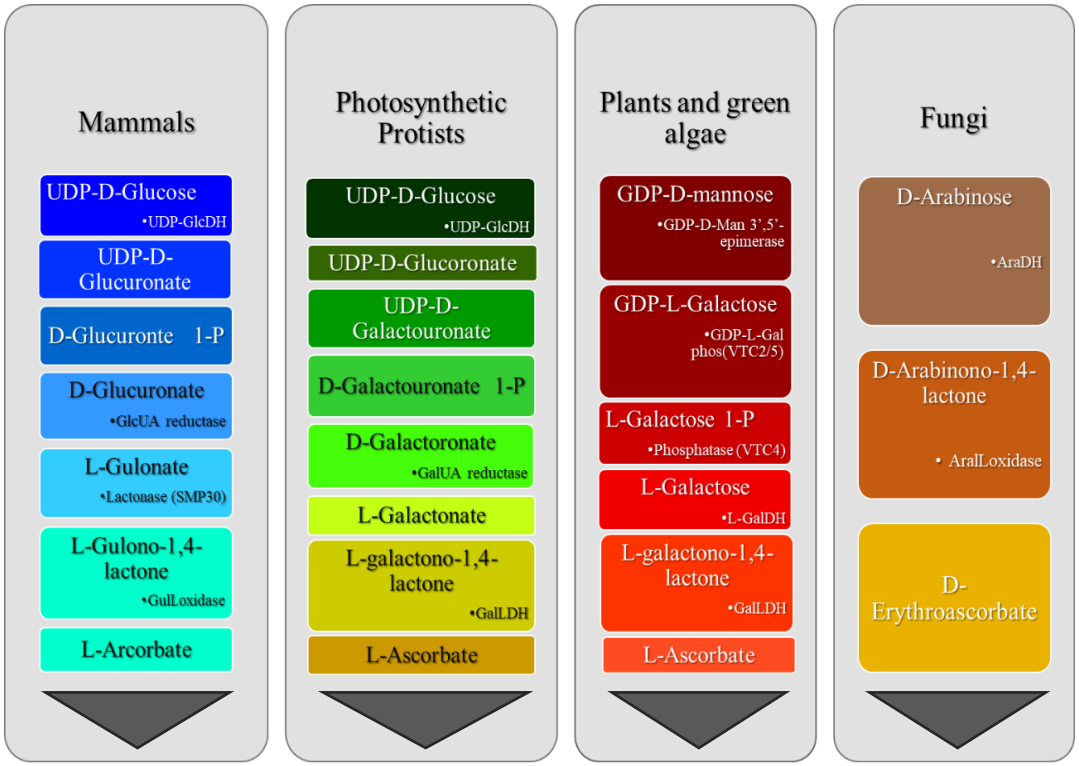


Figure 19 – Known ascorbic acid biosynthetic pathways. The last oxidation step of the distinct aldono-1,4-lactones to ascorbate is catalyzed by a FAD-linked oxidase or dehydrogenase (GULO, GALDH or ALO). It is to be noticed, that the photosynthetic protists seem to have some enzymatic components from mammal and plant pathways. As such, the described pathway for these species likely evolved from a secondary endosymbiosis event, regarding a non-photosynthetic ancestor and algae (Wheeler *et al.*, 2015). Presented figure and description were both adapted from Smirnoff., (2018).

The ascorbic acid group proteins are designated as SVCT1, SVCT2, SVCT3 and SVCT4 (de Koning and Diallinas, 2000; Yamamoto *et al.*, 2010). Curiously, from these four proteins, only SVCT1 and SVCT2 (the products of the *SLC23A1* and *SLC23A2* genes), are involved in ascorbic acid uptake and share a unique and specific conserved amino acid motif (SSSP) (Muñoz *et al.*, 2015; Kourkoulou, Pittis and Diallinas, 2018).

In humans, the SVCT1 transporter is mostly expressed in the epithelial tissues of several organs, such as the intestine, kidney and liver, while SVCT2 is expressed ubiquitously throughout the body (Lee *et al.*, 2006). Furthermore, it is known that the SVCT1 transporter contributes mainly to ascorbic acid uptake and therefore whole-body ascorbic acid level regulation, whereas the SVCT2 transporter is linked with specific responses to oxidative stress in the cells (Bürzle *et al.*, 2013).

Also, Kuo *et al.* (2004) demonstrated that the SVCT1 and SVCT2 transporters are likely to function and be expressed independently in mice (*Mus musculus*), since a lower expression of SVCT2 in heterozygous SVCT2 knockout individuals did not affect the expression of SVCT1 in the kidney and liver, condition that allowed for normal ascorbic acid levels in these organs.

Lastly, Kuo *et al.* (2004) also showed that the ascorbic acid levels in SVCT2-predominat organs, (brain or spleen), were lower in these mutant mice, which could indicate that this transporter is essential for the maintenance of ascorbic acid levels in tissues without notable presence of SVCT1.

In what concerns the *GULO* and *SVCT* genes, and after the results provided by Duque, (2018) it was possible to establish a relation between the evolution of the *Regucalcin* gene with the evolution of these two genes (figure 20).

Thus, in protostomia, for Porifera group, it is only possible to identify the *Regucalcin* gene. The Placozoa, controversially, just has the *GULO* and one *SVCT* genes. In Myxozoa, it is only possible to identify a *GULO* gene. Anthozoa, on the other hand, presents all the genes, having two identified *SVCT* genes.

In Hydrozoa group, it is only possible to identify one *SVCT* gene. Moreover, in Platyhelminthes, it seems that there is a missing link, due to the fact of presenting only the *Regucalcin* gene and one *SVCT* gene. The Mesozoa group is inferred to be lacking all the genes. Curiously, in Nematoda, both the *Regucalcin* and *GULO* genes are missing, yet they present three different *SVCT* genes. Priapulida only have *GULO*. In Tardigrada it is possible to observe an absence of all genes. The Chelicerata group without Araneae and Acari has the *Regucalcin* gene and one *STCV* gene. In Acari, it is possible to observe the presence of all genes, having two gene duplicates for *SVCT*. In Araneae it is possible to observe all genes. In Crustacea, only the *SVCT* gene is present. In the Collembola group, it is possible to observe that the *GULO* gene is missing and that the *Regucalcin* gene had initially three duplicates, where two of them were lost. The same thing happened for Phthiraptera, but only one of the three duplicates was lost. Considering the Hemiptera group, it was possible to identify three duplicates of the *Regucalcin* gene, and two of the *SVCT* gene and the *GULO* gene is missing. In the Coleoptera order, only four *Regucalcin* gene duplicates could be found. As for the Hymenoptera, two *Regucalcin* gene duplicates could be identified as well as one *SVCT* gene. In Diptera, eight *Regucalcin* gene duplicates were inferred, being one of them already described, the *Dca* gene in the Sophophora subgenus and also a *SVCT* gene was found. The Lepidoptera order, showed four inferred *Regucalcin* gene duplicates and one *SVCT* gene.

Afterwards, four *Regucalcin* gene duplicates are inferred for Brachiopoda, Annelida Gastropoda, Cephalopoda and Bivalvia. In the first case only gene 4 was kept, for *Regucalcin*, the *GULO* gene is present and two *SVCT* genes are observable. In Annelida both the second and fourth *Regucalcin* gene duplicates were kept, the *GULO* gene is present and two *SVCT* genes can be identified. In the case of Gastropoda the first, third and fourth *Regucalcin* gene duplicates were kept, the *GULO* gene is present and three *SVCT* duplicates were found. As for the Cephalopoda, the only kept duplicate for the *Regucalcin* gene is the fourth, and both *GULO* and the *SVCT* genes are missing. For Bivalvia it is exactly the same, with the exception that there is a *SVCT* gene.

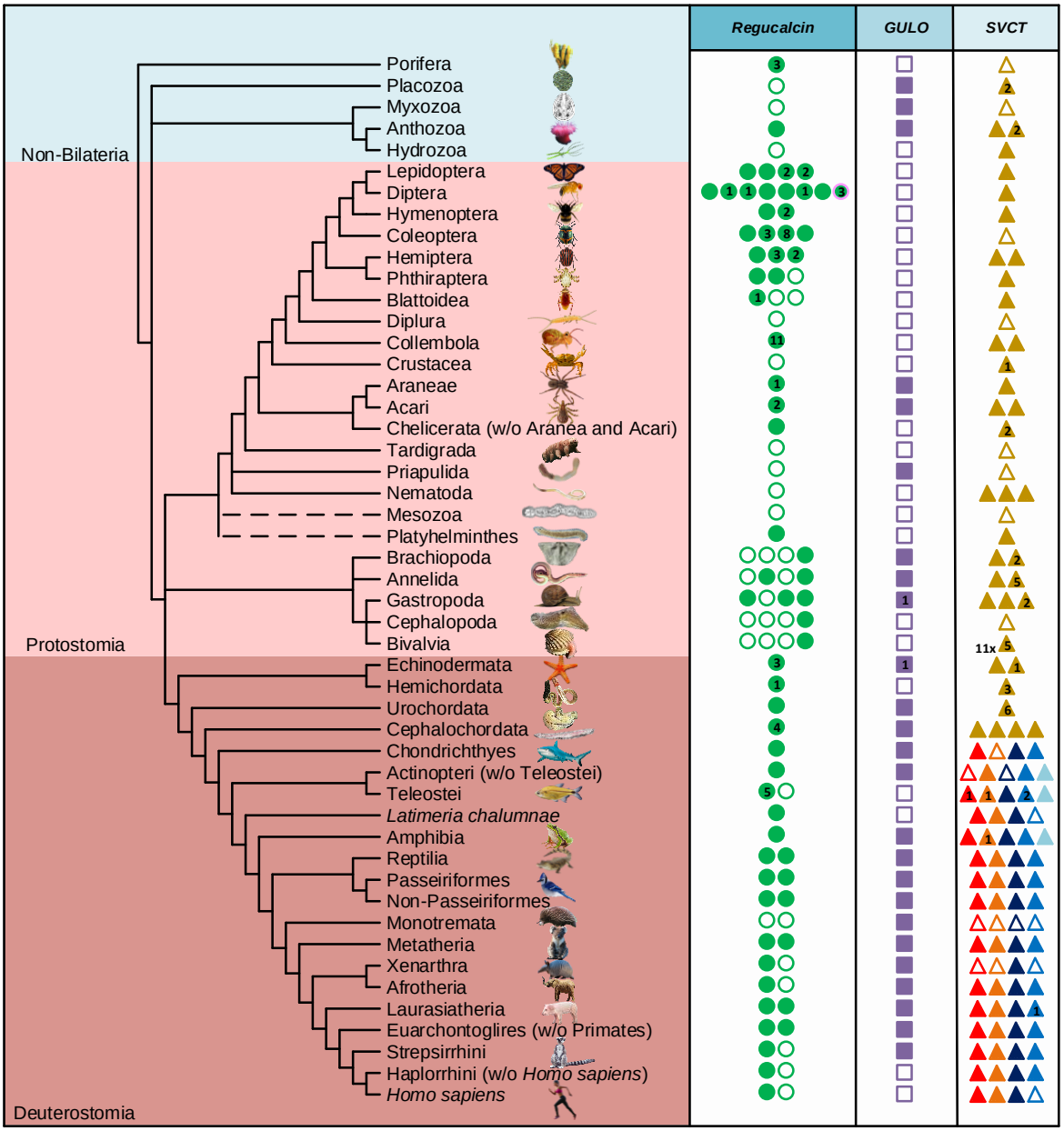


Figure 20 - Patterns of *Regucalcin*, *GULO* and *SVCT* gene presence/absence within the animal kingdom. Circles are representative of the *Regucalcin* gene, while squares and triangles represent the *GULO* and *SVCT* genes, respectively. Filled/empty figures are indicative of gene presence/absence evidence, and the number of figures for each gene indicates the identified duplicates within each group/species. Numbers highlighted in the figures represent the local duplications found within specific lineages. The pink colour outline represents gene duplications already described in current literature (*Dca* gene). The four known *SVCT* genes within the vertebrates lineage are represented in red (*SVCT*1), orange (*SVCT*2), dark blue (*SVCT*3) and blue (*SVCT*4), while the putative *SVCT*5 is highlighted in light blue. The remaining *SVCT* genes in basal deuterostomians, protostomians and non-bilaterians are represented in burnt yellow. The information regarding the *GULO* and *SVCT* genes was adapted from Duque, P. (2018).

Moreover, as for the basal deutorostomes in Echinodermata, all the genes are present, having *SVCT* two duplicates. In Hemichordata, only the *GULO* gene in missing. As for Urochordata, the presence of all genes is observable. Then, in Cephalochordata, it is possible to see all genes present, also, four identifiable duplicates of the *SVCT* gene.

From this evolutive point on, there are four *SVCT* gene duplicates already described (Kourkoulou, Pittis and Diallinas, 2018). In deuterostomes, more specifically the Chondrichthyes group, both the *Regucalcin* and *GULO* genes were present as well as *SVCT* 1, 2, 3 and 4. In Actinopteri (without Teleostei) all of the genes are present also with five *SVCT* duplicates. In Teleostei, there are two *Regucalcin* gene duplicates, where the second was lost (resulting from a whole genome duplication event), there is no *GULO* gene, but five *SVCT* gene duplicates were found. As for *Latimeria Chalumnae*, the *Regucalcin* gene is present, the *GULO* gene is absent and four *SVCT* gene duplicates have been found with a possible loss of *SVCT*4. In Amphibia all the genes are found, as well as five *SVCT* gene duplicates.

Furthermore, the groups of Reptilia, Passeriformes and Non-Passeriformes all have two *Regucalcin* gene duplicates, a *GULO* gene and the four *SVCT* gene duplicates. Monotremata probably lost its two *Regucalcin* gene duplicates, has the *GULO* gene and is inferred to have lost all the four *SVCT* gene duplicates. As for Metatheria, two *Regucalcin* gene duplicates are seen, the *GULO* gene is present as well as the four *SVCT* gene duplicates. In Xenarthra and Artotheria groups it is possible to identify two *Regucalcin* gene duplicates, where the second was posteriorly lost, both have the *GULO* gene and for the first one, only *SVCT*3 is identifiable and for the second group all the four described *SVCT* gene duplicates are observable. Both Laurasiatheria and Euarchontoglires (without primates) have two identified *Regucalcin* gene duplicates, the *GULO* gene is present, as well as all the four described *SVCT* gene duplicates.

Finally, both in Strepsirrhinii, Haplorrhini and in *Homo sapiens*, two *Regucalcin* gene duplicates can be seen, where the second was probably lost. From these three groups only in Strepsirrhinii the *GULO* gene is present. Additionally, both in Strepsirrhinii and in Haplorrhini all the four described *SVCT*s can be identified, in which *H. sapiens* just has *SVCT*1, *SVCT*2 and *SVCT*3.

In general, it was possible to observe that *Regucalcin* is duplicated more often in protostomians relatively to deuterostomian species. Contrary to this aspect, the deuterostomes seem to have more duplicates of the *STCV* gene (even if putative) than the protostomes. Nevertheless, the *GULO* gene does not appear to have been affected by ancestral duplications.

Additionally, it is possible to infer that in the cases where both *Regucalcin* gene (and most of its duplicates) and the *GULO* gene is likely lost in Protostomes, there is a tendency to frequently present one or more *SVCT* duplicates. Suggesting that in these groups there is only ascorbic acid

transportation of some sort. This happens in Hydrozoa, Blattoidea, Collembola, Nematoda and Bivalvia groups. Nevertheless, it is also possible that in these animal groups AA is synthetized by an unknown pathway as suggested for Drosophila (Henriques *et al.*, 2019) and Nematodes (Patananan *et al.*, 2015).

Another strange case seems to be happening in Deuterostomes, in which the Monotremata group is the only that lost both *Regucalcin* gene duplicates and all the four inferred *SVCT* genes. This group is represented by the only the egg-laying mammals, like the analysed species *Ornithorhynchus anatinus*. Nevertheless, they still have the *GULO* gene. Therefore, even if they're able to synthetize AA, they cannot transport it in its organism. Also by not presenting both of the *Regucalcin* gene duplicates, it is unclear, if there is another gene performing its attributed functions (calcium homeostasis, AA biogenesis, and so on) or if it had developed another way to do so. Therefore this could be an interesting matter for future work.

2. *Regucalcin* and *Dca* are essential genes

As depicted by Walsh, (1995), a new gene duplicate is able to rapidly evolve a new and significant function, by accumulating beneficial mutations, particularly in the species with large effective population sizes, such as the case of *Drosophila* (Kreitman and Comeron, 1999). These observations may perhaps explain why duplicate genes are as essential as singletons (Liang and Li, 2007; Liao and Zhang, 2007; Su and Gu, 2008; Makino, Hokamp and McLysaght, 2009). New genes are endlessly arising through various mechanisms, such as DNA-based duplication, retroposition, and de novo origination (Long *et al.*, 2003; Kaessmann, Vinckenbosch and Long, 2009).

This is in contrast to the view that essential genes are usually conserved and ancient (Miklos and Rubin, 1996; Krebs, Goldstein and Kilpatrick, 2018), while younger ones (normally existing in one or a few species) are considered "unnecessary" and attributed to perform relatively minor organismal functions (Wilson, Carlson and White, 1977; Miklos and Rubin, 1996; Krylov, 2003; Krebs, Goldstein and Kilpatrick, 2018).

Nevertheless, a study conducted in *Drosophila*, by Chen, Zhang and Long, (2010) demonstrated that 30% of newly arisen genes were essential for viability and that the proportion of genes perceived as essential was, in fact, similar in every evolutionary age group to the one that was examined.

Interestingly, most of the 59 young genes identified were highly expressed at the late larval stages or during metamorphosis, being that, the vast majority (47 of 59, 80%) of the young essential genes consistently showed lethality during pupation (Chen, Zhang and Long, 2010). The normal expression patterns of new genes was disrupted and it was seen that the development of the adult

organs was affected. About 50% of old genes are lethal during pupation, and the other half are lethal at earlier stages, because many early-stage developmental genes are conserved in accordance with Dietzl *et al.,* (2007). With these evidence, it is possible to speculate that new genes likely evolved essential functions in larval and pupal development, and that they normally regulate development in the pupal stage, with 10% or more regulating the development in the larval or even embryonic stages (Joppich *et al.*, 2009). As such, it seems that young essential genes have a tendency to play vital roles in middle or late stages of development, with a few cases in early stages.

Additionally, Chen, Zhang and Long, (2010) even showed that young essential genes had higher protein substitution rates than their parental genes, caused by either relaxation of functional constraint or positive selection, suggesting that, a new essential gene could arise from either an essential or a nonessential parent. Furthermore, these authors also show that either essential genes or nonessential genes can give rise to each type of gene (Chen, Zhang and Long, 2010).

Moreover, they hypothesize that a new gene might not become essential right after its appearance but may be integrated into a vital pathway by interacting with existing genes, being the interaction augmented by mutation and selection (Chen, Zhang and Long, 2010).

The results of the fly experiments (Table 1) showed that most flies in which the expression of both *Regucalcin* and *Dca* genes, were knockdowned (Wt), were not able to reach the adult stage, meaning that both of these genes are essential.

This was surprising given the fact that *Dca* is described as a Drosophila cold acclimation gene and that the flies were reared at 25ºC. So possibly either this *Regucalcin* gene duplicate assumes more important functions or by being a duplicate, the RNAi of the *Dca* fly strain is also allowing the knockdown of the *Regucalcin*, but it cannot be proven by mRNA extraction, due to its lethality.

Table 1 - Counting of *D. melanogaster* crossings.

Crossing	♀ Wt	♀ Curly	♂ Wt	♂ Curly	X^2 ♀	X^2 ♂
RGN ♂ **x Gal4** ♀	40	445	11	454	338,19 (P< 0,001)	422,04 (P< 0,001)
RGN ♀ **x Gal4** ♂	81	407	3	355	217,78 (P< 0,001)	346,10 (P< 0,001)
Dca ♂ **x Gal4** ♀	70	324	16	303	163,65 (P< 0,001)	258,21 (P< 0,001)
Dca ♀ **x Gal4** ♂	56	194	25	159	76,176 (P< 0,001)	97,59 (P< 0,001)

3. Interaction domain and its proximity to the protein's lid

In deuterostomians, Regucalcin is important for several biological processes, being one of them the calcium homeostasis (Laurentino *et al.*, 2012). Nevertheless, this protein still remains poorly

characterized in protostomians and non-bilaterians. Therefore, to infer a possible role of Regucalcin in calcium homeostasis in protostomians, several interactome analysis were performed concerning *D. melanogaster*, *Homo Sapiens* and later *C. elegans* (where there was no *Regucalcin* gene found).

Firstly, all *Regucalcin* interacting proteins in *D. melanogaster* were gathered (table 2). After this, the calcium interacting proteins of Homo sapiens were also observed (table 3) trying to establish possible homologies between these two organisms that can be seen in tables 4 and 5. Several orthologous genes were found, suggesting that the Regucalcin protein network is conserved among Protostomians and deuterostomians.

Resulting from this *D. melanogaster/H. sapiens* comparison and after an extensive analysis of the *D. melanogaster* interactome, 10 proteins were selected, given their described function related to calcium homeostasis (Chloride intracellular channel, Annexin B10, FK506-binding protein 14, supercoiling factor, Calreticulin, Translationally controlled tumor protein and Seipin), present in table 5. As well as for other functions, such as, ascorbic acid biosynthesis (Glutathione S transferase), even though the fact that humans do not synthetize it, that can be observed in table 5. And female pupae development (CG11267 and CG2862) (table 5). Afterwards, for all these proteins, a 3D structure prediction was performed as well as the inference of putative active and passive amino acid residues. These active and passive docking sites were posteriorly assembled in the molecular protein structure and the other in which the interaction levels were stronger (meaning that all the selected proteins overlapped were strongly docked for that specific amino acid) were marked in the structure of Regucalcin protein of *D. melanogaster* (figure 21).

Table 2 - Regucalcin interacting proteins in *D. melanogaster*

Gene ID	Name	Function
31185	Phosphogluconate dehydrogenase	Phosphogluconate dehydrogenase (decarboxylating) activity; NADP binding.
31359	ciboulot	Actin binding; actin monomer binding.
31760	Thioredoxin reductase-1	Antioxidant activity; protein homodimerization activity; electron transfer activity; thioredoxin-disulfide reductase activity; glutathione-disulfide reductase activity; flavin adenine dinucleotide binding.
32045	Heat shock protein 60A	Unfolded protein binding; ATP binding.
32499	CG8117	Nucleic acid binding; zinc ion binding.
32595	Cyclophilin 1	Peptidyl-prolyl cis-trans isomerase activity; cyclosporin A binding; unfolded protein binding.

32789	scully	Estradiol 17-beta-dehydrogenase activity; testosterone dehydrogenase (NAD+) activity; steroid dehydrogenase activity; acetoacetyl-CoA reductase activity; ribonuclease P activity; 7-beta-hydroxysteroid dehydrogenase (NADP+) activity.
33202	Stress induced phosphoprotein 1	Hsp90 protein binding.
33351	Enolase	Magnesium ion binding; phosphopyruvate hydratase activity
33397	Ubiquitin carboxy-terminal hydrolase	Thiol-dependent ubiquitin-specific protease activity.
33461	Phosphoglycerate kinase	ATP binding; ADP binding; phosphoglycerate kinase activity.
34264	GDP dissociation inhibitor	GDP-dissociation inhibitor activity; Rab GDP-dissociation inhibitor activity.
34363	eEF1delta	Elongation factor 1 beta central acidic region, eukaryote; Translation elongation factor EF1B, beta/delta chains, conserved site; Translation elongation factor EF1B, beta/delta subunit, guanine nucleotide exchange domain; Translation elongation factor EF1B/ribosomal protein S6; Translation elongation factor eEF-1beta-like superfamily.
34529	Deoxyuridine triphosphatase	Magnesium ion binding; dUTP diphosphatase activity.
34573	Discs large 5	Border follicle cell migration; antimicrobial humoral response.
36040	TER94	Protein binding; ATP binding; polyubiquitin modification-dependent protein binding; ATPase activity.
37290	Proliferating cell nuclear antigen	DNA binding; DNA polymerase processivity factor activity.
39424	CG10638	Alditol:NADP+ 1-oxidoreductase activity; oxidoreductase activity; alcohol dehydrogenase (NADP+) activity. It is involved in the biological process described with: oxidation-reduction process.
42783	CG10184	L-allo-threonine aldolase activity.

Table 3 - Regucalcin interacting proteins in *Homo sapiens*

Gene ID	Name	Function
3094	histidine triad nucleotide binding protein 1	Encodes a protein that hydrolyzes purine nucleotide phosphoramidates substrates, including AMP-morpholidate, AMP-N-alanine methyl ester, AMP-alpha-acetyl lysine methyl ester, and AMP-NH2. The encoded protein interacts with these substrates via a histidine triad motif. This gene is considered a tumor suppressor gene. In addition, mutations in this gene can cause autosomal recessive neuromyotonia and axonal neuropathy.
3336	heat shock protein family E (Hsp10) member 1	Encodes a major heat shock protein which functions as a chaperonin. Its structure consists of a heptameric ring which binds to another heat shock

		protein in order to form a symmetric, functional heterodimer which enhances protein folding in an ATP-dependent manner.
6647	superoxide dismutase 1	Binds copper and zinc ions and is one of two isozymes responsible for destroying free superoxide radicals in the body. The encoded isozyme is a soluble cytoplasmic protein, acting as a homodimer to convert naturally-occurring but harmful superoxide radicals to molecular oxygen and hydrogen peroxide. The other isozyme is a mitochondrial protein. Mutations in this gene have been implicated as causes of familial amyotrophic lateral sclerosis.
7067	thyroid hormone receptor alpha	Nuclear hormone receptor for triiodothyronine. It is one of the several receptors for thyroid hormone, and has been shown to mediate the biological activities of thyroid hormone.
9948	WD repeat domain 1	Protein containing 9 WD repeats. WD repeats are approximately 30- to 40-amino acid domains containing several conserved residues, mostly including a trp-asp at the C-terminal end. WD domains are involved in protein-protein interactions. The encoded protein may help induce the disassembly of actin filaments
9975	nuclear receptor subfamily 1 group D member 2	Nuclear hormone receptor family, specifically the NR1 subfamily of receptors. The encoded protein functions as a transcriptional repressor and may play a role in circadian rhythms and carbohydrate and lipid metabolism. Alternatively spliced transcript variants have been described.
51686	ornithine decarboxylase antizyme 3	Ornithine decarboxylase antizyme family, which plays a role in cell growth and proliferation by regulating intracellular polyamine levels. Expression of antizymes requires +1 ribosomal frameshifting, which is enhanced by high levels of polyamines. Antizymes in turn bind to and inhibit ornithine decarboxylase (ODC), the key enzyme in polyamine biosynthesis; thus, completing the auto-regulatory circuit.
151246	shugoshin 2	Biased expression in testis (RPKM 14.8), lymph node (RPKM 3.1) and 10 other tissues

Table 4 – Resulting interactome of *H. sapiens* in *D. melanogaster* (http://evoppi.sing-group.org/results/table/distinct/e6af5985-f197-44ff-b58b-98d397a421d5), information provided by *Flybase* and in https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl

Gene ID	Name	Function
39476	CG11267	GroES chaperonin superfamily; GroES-like superfamily. Chaperone binding; unfolded protein binding; metal ion binding. It is involved in the chaperone cofactor-dependent protein refolding. The phenotypic classes of alleles include: lethal - all die during larval stage; partially lethal - majority die; viable; visible; some die during pupal stage. Peak expression observed at stages throughout embryogenesis, during early larval stages, during late pupal stages, in adult female stages.
39251	Superoxide dismutase 1	Antioxidant activity; zinc ion binding; superoxide dismutase activity; superoxide dismutase copper chaperone activity; copper ion binding; protein homodimerization activity. Involved in age-dependent response to oxidative stress; oxidation-reduction process; regulation of autophagy of mitochondrion; regulation of terminal button organization; response to oxidative stress; determination of adult lifespan; aging; removal of superoxide radicals. The phenotypes of these alleles manifest in: larval segment; adult central nervous system; photoreceptor cell; developing material anatomical entity; female germline stem cell; adult dorsal vessel; sensory system neuron; adult circulatory system; central nervous system; germline cyst.
39999	Ecdysone-induced protein 75B	Binding (principal function); transcription regulator activity; molecular transducer activity; heterocyclic compound binding; nucleic acid binding; DNA-binding transcription factor activity, RNA polymerase II-specific; DNA-binding transcription factor activity; transcription factor activity, direct ligand regulated sequence-specific DNA binding; protein binding; steroid hormone receptor activity; transition metal ion binding; transcription coactivator activity.
39505	flare	Actin binding; actin filament binding. Sarcomere organization; establishment of planar polarity; actin filament depolymerization; positive regulation of actin filament depolymerization; imaginal disc-derived wing hair organization; negative regulation of Arp2/3 complex-mediated actin nucleation; locomotion; border follicle cell migration; regulation of actin filament depolymerization; regulation of actin filament polymerization.
40345	Ecdysone-induced protein 78C	Steroid hormone receptor activity; DNA-binding transcription factor activity; zinc ion binding; sequence-specific DNA binding. It is involved in the biological process described with: negative regulation of transcription, DNA-templated. 52 alleles are reported.
36307	Ornithine decarboxylase antizyme	Ornithine decarboxylase inhibitor activity. It is involved in the biological process described with: positive regulation of protein catabolic process.

Table 5 - Resulting interactome of *D. melanogaster* against *H. sapiens* (http://evoppi.sing-group.org/results/table/distinct/f2df2651-d79c-4ceb-afec-704fad0be042).

Gene ID	Name	Function
31185	Phosphogluconate dehydrogenase	NADP binding; phosphogluconate dehydrogenase (decarboxylating) activity. It is involved in the biological process described with: oxidation-reduction process; pentose-phosphate shunt; glucose homeostasis.
31245	Seipin	It is involved in the biological process described with 13 unique terms, many of which group under: lipid biosynthetic process; cellular lipid catabolic process; endoplasmic reticulum calcium ion homeostasis; regulation of plasma membrane bounded cell projection organization; response to starvation; positive regulation of lipid storage; calcium ion homeostasis; multicellular organism development; fatty acid metabolic process; regulation of cellular metabolic process; regulation of biosynthetic process; response to nutrient levels. It is involved in the biological process described with 13 unique terms, many of which group under: lipid biosynthetic process; cellular lipid catabolic process; endoplasmic reticulum calcium ion homeostasis; regulation of plasma membrane bounded cell projection organization; response to starvation; positive regulation of lipid storage; calcium ion homeostasis; multicellular organism development; fatty acid metabolic process; regulation of cellular metabolic process; regulation of biosynthetic process; response to nutrient levels.
31359	ciboulot	Actin binding; actin monomer binding. It is involved in the biological process described with: brain development; actin filament organization; larval central nervous system remodeling.
31760	Thioredoxin reductase-1	FAD/NAD(P)-binding domain; FAD/NAD(P)-binding domain superfamily; FAD/NAD-linked reductase, dimerisation domain superfamily; Pyridine nucleotide-disulphide oxidoreductase, class I; Pyridine nucleotide-disulphide oxidoreductase, class I, active site; Pyridine nucleotide-disulphide oxidoreductase, dimerisation domain; Thioredoxin/glutathione reductase selenoprotein. Glutathione-disulfide reductase activity; thioredoxin-disulfide reductase activity; electron transfer activity; protein homodimerization activity; antioxidant activity; flavin adenine dinucleotide binding. It is involved with determination of adult lifespan; response to hypoxia; cell redox homeostasis; oxidation-reduction process.

31816	Moesin	Cytoskeletal protein binding; phosphatidylinositol-4,5-bisphosphate binding; protein binding; microtubule binding; actin binding. It is involved in actin filament-based process; establishment or maintenance of bipolar cell polarity; establishment or maintenance of apical/basal cell polarity; establishment or maintenance of polarity of larval imaginal disc epithelium; male courtship behavior, veined wing extension; regulation of membrane potential in photoreceptor cell; plasma membrane bounded cell projection morphogenesis; cortical microtubule organization; establishment of localization; establishment of localization in cell; positive regulation of mitotic nuclear division; branching involved in open tracheal system development; neuron projection development; regulation of localization; cell proliferation.
32045	Heat shock protein 60A	Unfolded protein binding; ATP binding. It is involved in apoptotic mitochondrial changes; protein targeting to mitochondrion; protein refolding; mitochondrion organization; 'de novo' protein folding; protein import into mitochondrial intermembrane space; cellular response to heat.
32349	Chloride intracellular channel	Calcium ion binding; lipid binding; glutathione peroxidase activity; chloride channel activity. Involved in negative gravitaxis; response to alcohol; chloride transport.
32499	CG8117	Protein features are: Transcription elongation factor S-II, central domain; Transcription elongation factor S-II, central domain superfamily; Zinc finger, TFIIS-type. Zinc ion binding; nucleic acid binding. It is involved in transcription, DNA-templated. The phenotypes of these alleles manifest in: trichogen cell; ganglion mother cell; mesothoracic tergum; embryonic/larval neuroblast. The phenotypic classes of alleles include: partially lethal - majority die; visible; viable; some die during pupal stage; neuroanatomy defective; lethal.
32522	GTPase regulator associated with FAK	Phospholipid binding; GTPase activator activity; ubiquitin-dependent protein binding. Involved with: positive regulation of receptor internalization; negative regulation of epidermal growth factor receptor signaling pathway; Rho protein signal transduction.
32584	Proteasome α4 subunit	Endopeptidase activity; threonine-type endopeptidase activity. It is involved in the biological process described with: proteasomal ubiquitin-independent protein catabolic process; proteasomal protein catabolic process; proteasome-mediated ubiquitin-dependent protein catabolic process.
32595	Cyclophilin 1	Peptidyl-prolyl cis-trans isomerase activity; unfolded protein binding. It is involved with: protein peptidyl-prolyl isomerization; protein folding.

32789	scully	Acetoacetyl-CoA reductase activity; 7-beta-hydroxysteroid dehydrogenase (NADP+) activity; estradiol 17-beta-dehydrogenase activity; testosterone dehydrogenase (NAD+) activity; ribonuclease P activity; steroid dehydrogenase activity. Involved with: ecdysone metabolic process; acyl-CoA metabolic process; steroid metabolic process; androgen metabolic process; mitochondrial tRNA processing; estrogen metabolic process; fatty acid metabolic process.
33019	Annexin B10	Calcium ion binding; actin binding; calcium-dependent phospholipid binding.
33202	Stress induced phosphoprotein 1	Hsp90 protein binding.
33351	Enolase	Magnesium ion binding; phosphopyruvate hydratase activity. Involved in the glycolytic process; glucose homeostasis. The phenotype of these alleles manifest in: pupa. The phenotypic classes of alleles include: phenotype; visible; lethal - all die before end of P-stage; flightless; lethal; fertile; increased mortality during development.
33076	CG1532	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase; Vicinal oxygen chelate (VOC) domain
33078	Nuclear transport factor-2	Protein transporter activity.
33397	Ubiquitin carboxy-terminal hydrolase	Thiol-dependent ubiquitin-specific protease activity. It is involved in the biological process described with: ubiquitin-dependent protein catabolic process; protein deubiquitination.
33461	Phosphoglycerate kinase	ATP binding; ADP binding; phosphoglycerate kinase activity. Involved in myoblast fusion; somatic muscle development; muscle cell cellular homeostasis; gluconeogenesis; chemical synaptic transmission; phosphorylation; glycolytic process.
33471	CG2862	Catalytic activity. Peak expression observed within 00-06 hour embryonic stages, during early larval stages, in adult female stages.
33487	Ribosomal protein S21	Ribosome binding; structural constituent of ribosome. It is involved in regulation of cell proliferation; translation; endonucleolytic cleavage to generate mature 3'-end of SSU-rRNA from (SSU-rRNA, 5.8S rRNA, LSU-rRNA); endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA); lymph gland development; cytoplasmic translation.
33981	smt3	Protein tag; protein binding. Involved in cell cycle; cell projection organization; cell part morphogenesis; regulation of MAP kinase activity; response to chemical; establishment of protein localization to organelle; regulation of protein localization to cell periphery; nervous system development; regulation of phosphorus metabolic process; response to transforming growth factor beta; establishment of protein localization; microtubule-based process; positive

		regulation of innate immune response; regulation of protein localization to membrane; lipid homeostasis.
34170	A kinase anchor protein 200	Actin filament binding; Notch binding; protein kinase A regulatory subunit binding; protein binding. Involved in positive regulation of Notch signaling pathway; regulation of actin cytoskeleton organization; regulation of actin cytoskeleton reorganization; negative regulation of Ras protein signal transduction; regulation of establishment of planar polarity; circadian rhythm; cellular response to ethanol; positive regulation of protein localization to membrane; behavioral response to ethanol; germ cell development.
34264	GDP dissociation inhibitor	GDP-dissociation inhibitor activity; Rab GDP-dissociation inhibitor activity. It is involved in protein transport; vesicle-mediated transport; small GTPase mediated signal transduction; neurotransmitter secretion.
34281	thioredoxin-2	Disulfide oxidoreductase activity; protein disulfide oxidoreductase activity. It is involved in cell redox homeostasis; response to oxidative stress; glycerol ether metabolic process; defense response to fungus; determination of adult lifespan.
34313	yippee interacting protein 2	Acetyl-CoA C-acetyltransferase activity; acetyl-CoA C-acyltransferase activity. Involved in fatty acid beta-oxidation.
34315	CG4548	Dodecenoyl-CoA delta-isomerase activity; enoyl-CoA hydratase activity. It is involved in the biological process described with: fatty acid beta-oxidation.
34363	eukaryotic translation elongation factor 1 delta	Translation elongation factor activity. Involved in translational elongation.
34500	porin	Voltage-gated anion channel activity. It is involved in the biological process described with: sperm mitochondrion organization; anion transmembrane transport; ion transport; mitochondrial transport; mitochondrion organization; sperm individualization; phototransduction.
34529	Deoxyuridine triphosphatase	dUTP diphosphatase activity; magnesium ion binding. Involved in dUMP biosynthetic process; dUTP catabolic process; dUTP metabolic process.
34573	Discs large 5	Involved in border follicle cell migration; antimicrobial humoral response.
34953	Endonuclease G inhibitor	Hydrolase activity, acting on acid anhydrides; ribonuclease inhibitor activity; telomeric DNA binding; double-stranded DNA binding. Involved in positive regulation of Notch signaling pathway; negative regulation of endoribonuclease activity; double-strand break repair via nonhomologous end joining; telomere maintenance; motor neuron axon guidance; cellular response to gamma radiation; cellular response to X-ray.

35192	lethal (2) 37Cb	ATP-dependent 3'-5' RNA helicase activity; RNA binding; ATP-dependent RNA helicase activity. It is involved in the biological process described with: mRNA splicing, via spliceosome.
35418	Decondensation factor 31	Histone binding. It is involved in the biological process described with: nucleosome assembly; chromatin organization.
35422	eukaryotic translation elongation factor 2	GTP binding; GTPase activity; translation elongation factor activity. It is involved in the biological process described with: translational elongation.
35584	Eb1	Microtubule binding; protein binding; myosin binding; microtubule plus-end binding. It is involved in behavior; adult locomotory behavior; metabolic process; sensory organ development; sister chromatid segregation; sensory perception of mechanical stimulus; pseudocleavage involved in syncytial blastoderm formation; response to wounding; anatomical structure development; developmental process.
35671	myo-inositol-1-phosphate synthase	Inositol-3-phosphate synthase activity. It is involved in inositol biosynthetic process; phospholipid biosynthetic process.
35728	Glyceraldehyde 3 phosphate dehydrogenase 1	Aldehyde or oxo oxidoreductases with NAD or NADP as acceptor, include dehydrogenases that oxidize an aldehyde or ketone (oxo) group with the reduction of NAD or NADP.
36040	TER94	Necessary for the fragmentation of Golgi stacks during mitosis and for their reassembly after mitosis. Involved in the formation of the transitional endoplasmic reticulum (tER).
36059	14-3-3ζ	Protein domain specific binding; transcription factor binding; protein homodimerization activity; protein kinase C inhibitor activity; protein heterodimerization activity; protein binding. It is involved in behavior; learning or memory; establishment or maintenance of cytoskeleton polarity; regulation of DNA-binding transcription factor activity; response to temperature stimulus; chromosome segregation; positive regulation of Ras protein signal transduction; transmembrane receptor protein tyrosine kinase signaling pathway; cellular nitrogen compound biosynthetic process; regulation of protein stability; protein folding.
36098	Peroxiredoxin 2540-2	Peroxidase activity; peroxiredoxin activity; protein binding. It is involved in cellular response to oxidative stress; cell redox homeostasis; response to oxidative stress; hydrogen peroxide catabolic process; oxidation-reduction process.
36270	Endoplasmic reticulum p60	Intramolecular oxidoreductases transposing S-S bonds catalyze the oxidation of one part of a molecule and the concurrent reduction of another part of the same molecule, and one or more sulfur-sulfur bonds in the molecule are rearranged.
36409	Nascent polypeptide associated complex protein alpha subunit	Protein binding.

36718	Kinesin heavy chain 73	Protein homodimerization activity; microtubule binding; ATPase activity; microtubule motor activity; ATP binding. Involved in microtubule-based movement; regulation of synapse structure or activity; establishment of spindle orientation.
36981	eukaryotic translation initiation factor 3 subunit b	Protein binding; translation initiation factor activity; mRNA binding; translation initiation factor binding. It is involved in ovarian follicle cell development; regulation of translational initiation; translational initiation.
37111	Glutathione S transferase E6	Glutathione transferase activity. It is involved in glutathione metabolic process.
37214	FK506-binding protein 12kD	Peptidyl-prolyl cis-trans isomerase activity. It is involved in chaperone-mediated protein folding; protein peptidyl-prolyl isomerization.
37290	Proliferating cell nuclear antigen	DNA binding; DNA polymerase processivity factor activity. It is involved in mitotic spindle organization; eggshell chorion gene amplification; nucleotide-excision repair; leading strand elongation; antimicrobial humoral response; DNA-dependent DNA replication; DNA replication; mismatch repair; regulation of DNA replication; translesion synthesis.
37449	FK506-binding protein 14	Calcium ion binding. It is involved in the biological process described with: imaginal disc-derived wing margin morphogenesis; muscle cell cellular homeostasis; imaginal disc development; regulation of Notch signaling pathway; chaeta development.
37467	Rae1	RNA binding; ubiquitin binding; protein binding. It is involved in spermatogenesis; male meiosis I; regulation of G1/S transition of mitotic cell cycle; negative regulation of synaptic growth at neuromuscular junction; regulation of autophagy; transcription-dependent tethering of RNA polymerase II gene DNA at nuclear periphery; positive regulation of gene expression; mitotic cell cycle.
37591	CG2852	Peptidyl-prolyl cis-trans isomerase activity; cyclosporin A binding; unfolded protein binding. It is involved in multicellular organism reproduction; protein folding; protein peptidyl-prolyl isomerization.
37617	bellwether	ATP binding; proton-transporting ATP synthase activity, rotational mechanism; ADP binding; ATPase activity. It is involved in the biological process described with: ATP metabolic process; response to oxidative stress; regulation of choline O-acetyltransferase activity; lipid storage; electron transport chain; ATP biosynthetic process; ATP synthesis coupled proton transport. 32 alleles are reported.
37834	Adenylate kinase 2	Adenylate kinase activity; ATP binding. It is involved in ADP biosynthetic process.
37846	eukaryotic translation elongation factor 5	Translation elongation factor activity; ribosome binding. It is involved in the biological process described with: positive

		regulation of translational termination; translational frameshifting; regulation of apoptotic process; multicellular organism reproduction; positive regulation of translational elongation.
38001	zipper	Myosin II or non-muscle myosin. motor protein, crucial functions in motility, cytokinesis, dorsal closure and cytoplasmic transport
38022	CG7049	Formylglycine-generating oxidase activity.
38145	supercoiling factor	Calcium ion binding. It is involved in chromatin organization; dosage compensation by hyperactivation of X chromosome; positive regulation of DNA topoisomerase (ATP-hydrolyzing) activity; DNA topological change; positive regulation of transcription, DNA-templated.
38301	CG8993	Disulfide oxidoreductase activity; protein disulfide oxidoreductase activity. It is involved in glycerol ether metabolic process; cell redox homeostasis.
38389	Heat shock protein 83	Disordered domain specific binding; unfolded protein binding; TPR domain binding; protein homodimerization activity; ATP binding; protein binding. It is involved in the response to abiotic stimulus; response to heat; regulation of cell proliferation; protein folding; NLS-bearing protein import into nucleus; cell cycle process; oocyte axis specification; regulation of biological quality; cellular response to heat; response to stimulus; regulation of biological process. 60 alleles are reported. The phenotypes of these alleles manifest in: larval head; histaminergic neuron; somatic muscle; centrosome; intracellular membrane-bounded organelle; appendage; imaginal tissue; adult metathoracic sensillum; Z disc; eye-antennal disc.
38413	PHGPx	Glutathione peroxidase activity; peroxidase activity. It is involved in oxidation-reduction process; response to oxidative stress; intestinal stem cell homeostasis; response to endoplasmic reticulum stress; response to lipid hydroperoxide.
38490	Chd64	Protein binding; actin binding; juvenile hormone response element binding. It is involved in juvenile hormone mediated signaling pathway; smooth muscle contraction.
38612	Cytochrome c1	Electron transporter, transferring electrons within CoQH2-cytochrome c reductase complex activity; heme binding. It is involved in mitochondrial electron transport, ubiquinol to cytochrome c; mitochondrial ATP synthesis coupled proton transport. 9 alleles are reported. The phenotypes of these alleles manifest in: trichogen cell; adult heart.
38820	Sh3β	SH3-binding, glutamic acid-rich protein; Thioredoxin-like superfamily.
38972	Glutathione S transferase O3	Glutathione dehydrogenase (ascorbate) activity; transferase activity, transferring sulfur-containing groups; glutathione transferase

		activity. Involved in glutathione metabolic process; oxidation-reduction process.
39251	Superoxide dismutase 1	Antioxidant activity; zinc ion binding; superoxide dismutase activity; superoxide dismutase copper chaperone activity; copper ion binding; protein homodimerization activity. It is involved in age-dependent response to oxidative stress; oxidation-reduction process; regulation of autophagy of mitochondrion; regulation of terminal button organization; response to oxidative stress; determination of adult lifespan; aging; removal of superoxide radicals.
39304	CG6084	Oxidoreductase activity; indanol dehydrogenase activity; alditol:NADP+ 1-oxidoreductase activity; alcohol dehydrogenase (NADP+) activity. It is involved in oxidation-reduction process.
39364	viral IAP-associated factor	It is involved in: regulation of cysteine-type endopeptidase activity involved in apoptotic process.
39424	CG10638	Oxidoreductase activity; alditol:NADP+ 1-oxidoreductase activity; alcohol dehydrogenase (NADP+) activity. It is involved in oxidation-reduction process.
39476	CG11267	Chaperone binding; unfolded protein binding; metal ion binding. It is involved in chaperone cofactor-dependent protein refolding.
39505	flare	Actin binding; actin filament binding. It is involved in sarcomere organization; establishment of planar polarity; actin filament depolymerization; positive regulation of actin filament depolymerization; imaginal disc-derived wing hair organization; negative regulation of Arp2/3 complex-mediated actin nucleation; locomotion; border follicle cell migration; regulation of actin filament depolymerization; regulation of actin filament polymerization.
39557	Hsc70Cb	Heat shock protein 70 family; Heat shock protein 70, conserved site; Heat shock protein 70kD, C-terminal domain superfamily; Heat shock protein 70kD, peptide-binding domain superfamily.
39651	Protein disulfide isomerase	Peptide disulfide oxidoreductase activity; protein disulfide isomerase activity; peptidyl-proline 4-dioxygenase activity. It is involved in regulation of oxidative stress-induced intrinsic apoptotic signaling pathway; cell redox homeostasis; protein folding; response to endoplasmic reticulum stress.
39826	failed axon connections	Axonogenesis; female germ-line stem cell population maintenance.
39933	Myoinhibiting peptide precursor	G-protein coupled receptor binding; signaling receptor binding; neuropeptide hormone activity. It is involved in negative regulation of juvenile hormone biosynthetic process; negative regulation of eating behavior; G-protein coupled receptor signaling pathway; neuropeptide signaling pathway.
40179	RhoGDI	Rho GDP-dissociation inhibitor activity; Rac GTPase binding. It is involved in Rho protein signal transduction.

41027	CG8036	Catalytic activity. It is involved in regulation of chromatin silencing.
41166	Calreticulin	Unfolded protein binding; calcium ion binding. It is involved in olfactory behavior; protein folding; sleep.
41173	p23	Hsp90 protein binding; chaperone binding; prostaglandin-E synthase activity. It is involved in protein folding; Golgi organization; chaperone-mediated protein complex assembly.
41341	Translationally controlled tumor protein	Guanyl-nucleotide exchange factor activity; calcium ion binding. It is involved with intra-S DNA damage checkpoint; positive regulation of multicellular organism growth; positive regulation of histone phosphorylation; cellular response to gamma radiation; double-strand break repair; mitotic G2 DNA damage checkpoint; regulation of stem cell differentiation; positive regulation of cell size.
42185	Malate dehydrogenase 2	Malate dehydrogenase activity; L-malate dehydrogenase activity. It is involved in pupal development; larval midgut cell programmed cell death; tricarboxylic acid cycle; positive regulation of programmed cell death; salivary gland histolysis; oxidation-reduction process; salivary gland cell autophagic cell death; activation of cysteine-type endopeptidase activity involved in apoptotic process; carbohydrate metabolic process; regulation of programmed cell death.
42186	14-3-3ε	Phosphoserine residue binding; protein heterodimerization activity; protein binding; protein domain specific binding; transcription factor binding. It is involved with regulation of growth; response to radiation; cell cycle process; multicellular organism aging; regulation of RNA metabolic process; organelle fission; neuron differentiation; regulation of small GTPase mediated signal transduction; response to abiotic stimulus; growth.
42783	CG10184	L-allo-threonine aldolase activity. It is involved in threonine catabolic process; glycine biosynthetic process.
43183	Aldolase 1	Fructose-bisphosphate aldolase activity. It is involved with mesoderm development; glucose homeostasis; glycolytic process.
43447	Phosphoglyceromutase 78	Phosphoglycerate mutase activity. It is involved in the biological process described with: glycolytic process; somatic muscle development; myoblast fusion.
43448	Stem-loop binding protein	Histone pre-mRNA stem-loop binding; mRNA binding; protein binding; RNA stem-loop binding. It is involved in mRNA 3'-end processing by stem-loop binding and cleavage; mRNA transport; mitotic chromosome condensation.
43560	Nucleoplasmin	chromatin binding; histone binding. It is involved in the biological process described with: chromatin remodeling; sperm chromatin decondensation.

43739	abnormal wing discs	Nucleoside diphosphate kinase activity; GTP binding; microtubule binding; ATP binding; kinase activity; magnesium ion binding. It is involved in biological regulation; tissue morphogenesis; cellular component organization or biogenesis; microtubule-based process; cellular component organization; purine nucleotide biosynthetic process; establishment or maintenance of cell polarity; mitotic cell cycle; pyrimidine nucleoside triphosphate biosynthetic process; cell junction organization; anatomical structure morphogenesis; cell cycle; tissue migration; macromolecule modification.
43429	CG11837	rRNA (adenine-N6,N6-)-dimethyltransferase activity. It is involved in the biological process described with: rRNA methylation.
45233	capulet	Actin binding; adenylate cyclase binding. Involved in compound eye development; visual system development; sensory organ development; regulation of catalytic activity; muscle cell cellular homeostasis; tube development; axonogenesis; chaeta morphogenesis; biological regulation; wing disc pattern formation; regulation of cellular component organization; regulation of lyase activity; pattern specification process.
45780	Proteasome α1 subunit	Some flies die during embryonic stage; lethal - all die before end of embryonic stage. Involved in proteasomal protein catabolic process; proteasome-mediated ubiquitin-dependent protein catabolic process; proteasomal ubiquitin-independent protein catabolic process.
48228	Deoxyribonuclease II	Endopeptidase activity; threonine-type endopeptidase activity.
3772232	fatty acid binding protein	Lipid binding. It is involved in long-term memory.

After the interactomes performed for the Regucalcin protein, in *Drosophila melanogaster* it was time to further investigate their docking sites to the Regucalcin molecule and its active and passive residues. Thus, in table 6, a calcium interaction in *D. melanogaster*, can be observed, as well as its docking sites. Some other studies were made, and are presented here, both on development (Female pupae) in *D. melanogaster* (table 7) and in the Ascorbic Acid pathway in *D. melanogaster* (table 8).

Table 6 – Results of the docking prediction analysis regarding calcium-related interactors in *D. melanogaster*

Gene name	Gene ID	Uniprot ID	Protein size	Active residues	Passive residues
Chloride intracellular channel	32349	Q9VY78	260	1,2,3,5,12,16,17,19,20,21,31,44,45,46,47,48,49,50,70,71,73,75,76,77,78,79,80,81,82,127,128,131,136,137,138,140,172	7,8,9,10,11,22,23,33,35,36,37,41,42,43,51,52,53,63,64,65,67,69,74,83,84,85,87,88,89,92,96,122,124,125,126,129,139,141,149,150,152,157,167,168,169,170,171
Annexin B10	33019	P22465	321	1,2,3,4,5,6,7,8,10,11,47,48,87,88,89,117,118,119,147,148,160,166,170,204,205,208,209,275,276,279,280,284	12,13,14,15,44,45,51,52,79,82,86,90,94,115,116,120,121,123,124,126,127,131,132,146,150,151,155,158,159,161,162,163,164,165,167,168,169,171,172,174,192,193,201,212,213,215,233,234,244,245,246,247,248,250,268,283,287,288,291,315,316,319,320
FK506-binding protein 14	37449	Q9V3V2	216	19,20,21,22,27,28,50,62,64,71,72,73,82,84,85,98,108,109,110,112,113,116,119,121,122,123,124,153,160,162,163,164,165,169,170,171,194,198,215,221,222,224,228,236,237,246,255,256,257,258,265,282,283,291,292,298,300,303,305,306,310,316,318,319	26,27,28,39,40,43,44,46,62,68,75,78,79,80,83,91,98,101,103,105,111,112,113,116,118,132,167,168,169,170
supercoiling factor	38145	Q9W0H8	329	1,2,3,4,5,7,8,11,12,13,15,16,17,18,19,20,22,23,24,26,27,32,36,42,43,44,46,48,49,57,58,61,112,113,114,319,323,324,325,326,327	6,10,25,28,29,30,31,33,34,35,37,38,39,40,41,45,47,50,51,52,53,54,55,56,59,62,65,76,79,83,89,90,93,95,99,100,103,107,109,110,111,115,116,119,122,129,130,131,139,142,143,146,147,150,172,180,211,215,217,225,228,254,314,315,317,318,321,322
Calreticulin	41166	P29413	406	1,2,3,4,5,6,8,9,11,12,13,14,15,16,18,19,20,25,26,28,29,35,80,81,82,83,84,241,338	30,31,33,34,36,37,61,62,85,87,88,89,90,115,119,238,239,240,242,243,265,266,267,271,309,

					310,335,336,337,340,341,342,345,346
Translationally controlled tumor protein	41341	Q9VGS2	172	1,2,3,5,12,16,17,19,20,21,31,70,71,73,75,76,77,78,79,80,81,82,127,128,131,136,137,138,140	7,8,9,10,11,22,23,33,35,36,37,42,44,63,64,65,67,69,74,83,84,85,87,88,89,92,96,122,124,125,126,129,139,141,149,150,152,157,167,168,169,170,171,172
Seipin	31245	Q9V3X4	370	1,2,3,4,6,7,8,9,11,12,13,14,17,24,26,27,28,29,31,32,49,51,52,54,55,58,59,62,63,66,67,69,71,75,94,141,144,145,147,166,171,172,173,174,175,176,177,178,179,180,181,182,183,184,185,186,187,188,189,190,191,192,195,240,241,242,244,245,246,247,249,250,251,253,254,255,257,258,259,261,262,263,264,265,266,269,270,305,315,316,317,319,322,324,325	16,19,34,35,36,38,40,41,42,43,44,46,47,50,74,76,79,90,92,95,97,99,105,106,113,115,116,123,124,125,127,135,136,137,138,139,142,143,153,156,157,158,159,162,164,165,168,170,193,194,196,202,217,223,224,227,228,230,238,239,275,278,281,282,286,287,289,290,292,293,300,301,302,303,304,328,330,331,332,333,335,336,338,339,346,347,369,370
Regucalcin	32164	Q76NR6	319	1,2,3,5,6,7,8,9,10,11,12,13,14,16,18,23,25,26,83,95,111,114,115,120,143,144,145,146,147,148,149,150,151,152,161,174,193,259,261,307,308	19,20,21,22,27,28,50,62,64,71,72,73,82,84,85,98,108,109,110,112,113,116,119,121,122,123,124,153,160,162,163,164,165,169,170,171,194,198,215,221,222,224,228,236,237,246,255,256,257,258,265,282,283,291,292,298,300,303,305,306,310,316,318,319

Table 7 - Results of the docking prediction analysis regarding developmental interactors (Female pupae) in *D. melanogaster*

Gene name	Gene ID	Uniprot ID	Protein size	Active residues	Passive residues
CG11267	39476	Q9VU35	103	1,2,3,4,5,6,7,9,11,12,14,23,24, 25,26,27,28,29,30,31,32,33,34, 35,36,37,38,46,50,52,53,56,57, 58,59,60,61,62,64,66,67,68,72, 74,75, 76,77,78,79,82,92,94,95,96,97,9 8,99,100,102,103	18,21,22,39,40,41,42,44, 54,55,63,69,70,80,81,84,85, 86,87,88,89,91,101
CG2862	33471	Q8STA5	150	2,3,4,5,7,8,12,13,14,15,16,17,18, 19,20,,23,24,51,52,54,55,56,57, 60,66,69,70,71,73,87,88,95,96, 98,99,100,102,103,106,117,118, 119,120,121,122,123,124,125,12 6,127,128,129,130,131,133,142, 143,144,145,146,147,148,149, 150	26,27,28,39,41,42,47,48,49, 50,58,59,61,81,83,85,86,89,90, 91,92,93,97,107,110,111, 112,115,116
Regucalcin	32164	Q76NR6	319	1,2,3,5,6,7,8,9,10,11,12,13,14, 16,18,23,25,26,83,95,111,114, 115,120,143,144,145,146,147, 148,149,150,151,152,161,174, 193,259,261,307,308	19,20,21,22,27,28,50,62,64,71, 72,73,82,84,85,98,108,109,110, 112,113,116,119,121,122,123, 124,153,160,162,163,164,165, 169,170,171,194,198,215,221, 222,224,228,236,237,246,255, 256,257,258,265,282,283,291, 292,298,300,303,305,306,310, 316,318,319

Table 8 - Results of the docking prediction analysis regarding ascorbic acid homeostasis interactors in *D. melanogaster*

Gene name	Gene ID	Uniprot ID	Protein size	Active residues	Passive residues
Glutathione S transferase	38972	Q9VSL2	241	1,2,3,4,5,7,31,32,48,50,54,57,62, 63,66,67,68,69,83,84,85,86,88,89, 92,93,105,107,108,109,111,112,113, 115,116,118,119,122,126,238, 239,240,241	6,8,9,10,11,12,13,14,15,18,21,41, 46,47,55,56,59,64,77,81,82,95,96, 97,98,102,104,106,110,120,123,125, 129,130,155,161,166,216,218, 219,222,224,225,228,230,234, 235,236,237
Regucalcin	32164	Q76NR6	319	1,2,3,5,6,7,8,9,10,11,12,13,14, 16,18,23,25,26,83,95,111,114, 115,120,143,144,145,146,147, 148,149,150,151,152,161,174, 193,259,261,307,308	19,20,21,22,27,28,50,62,64,71,72, 73,82,84,85,98,108,109,110,112, 113,116,119,121,122,123,124,153, 160,162,163,164,165,169,170,171, 194,198,215,221,222,224,228,236, 237,246,255,256,257,258,265, 282,283,291,292,298,300,303, 305,306,310,316,318,319

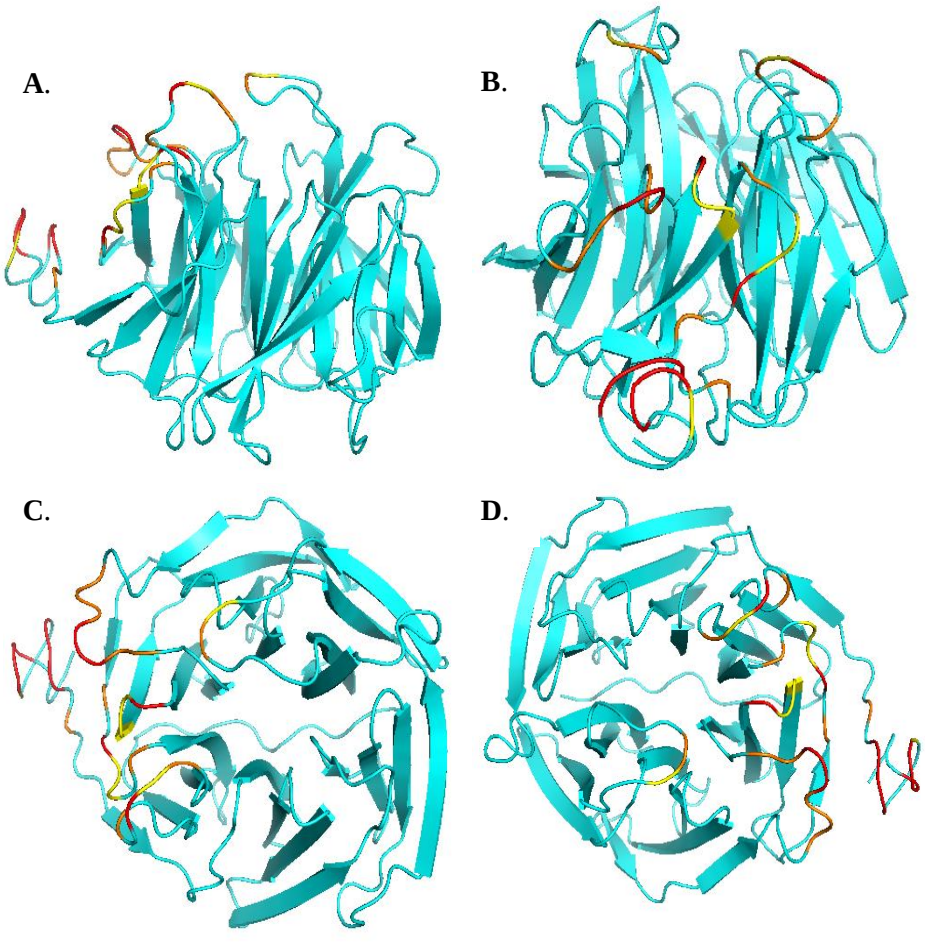


Figure 21- Putative docking sites of the *D. melanogaster* Regucalcin protein. The red colour indicates maximum docking hit (total of 10 proteins), orange represents nine docking hits and yellow, eight docking hits.

In figure 21, it is possible to observe the three different colour hot intensities over the blue Regucalcin structure, these represents the number of proteins, out of the ten analysed, inferred to make a contact in that precise location. Surprisingly, it is clear that all the selected proteins interact with Regucalcin in a confined location of the protein structure. An example of this interaction can be found in supplementary material in chapter 3. This observation is important, given the fact that little is known about how exactly this molecule interacts with other proteins, or if it has several interaction points. What these findings indicate is that, it is likely that proteins (of different sorts) interact predominantly in one spot of the Regucalcin protein molecule, and so it is possible to suggest that maybe this spot is in fact an interaction domain.

Defined by (Aizawa *et al.*, 2013) on the top of the molecule it is possible to observe what the authors described both in humans and in mice as the lid loop, which can be observed in figure 22.

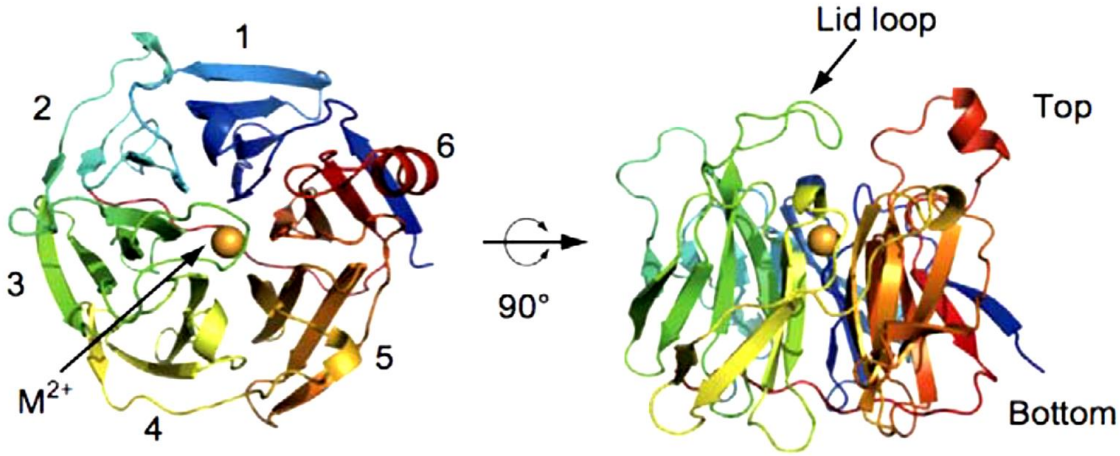


Figure 22 - Overall structure of mouse SMP30/GNL. The structure is shown as a rainbow coloured cartoon with N-terminus = blue and Cterminus = red. The divalent metal ion (labeled as M2+) located at the center of the structure is shown as an orange sphere. Presented figure and description were both adapted from Aizawa *et al.* (2013).

This lid is located in the superior part of the Regucalcin molecule and as it can be seen in figure 22, it exists in more species than just the two described. By looking to figure 23, it is possible to infer that this lid shares similarities with the *D. melanogaster* structure. In addition, that these putative docking sites, that constitute the proposed interaction domain are close to the protein lid or even overlapping it.

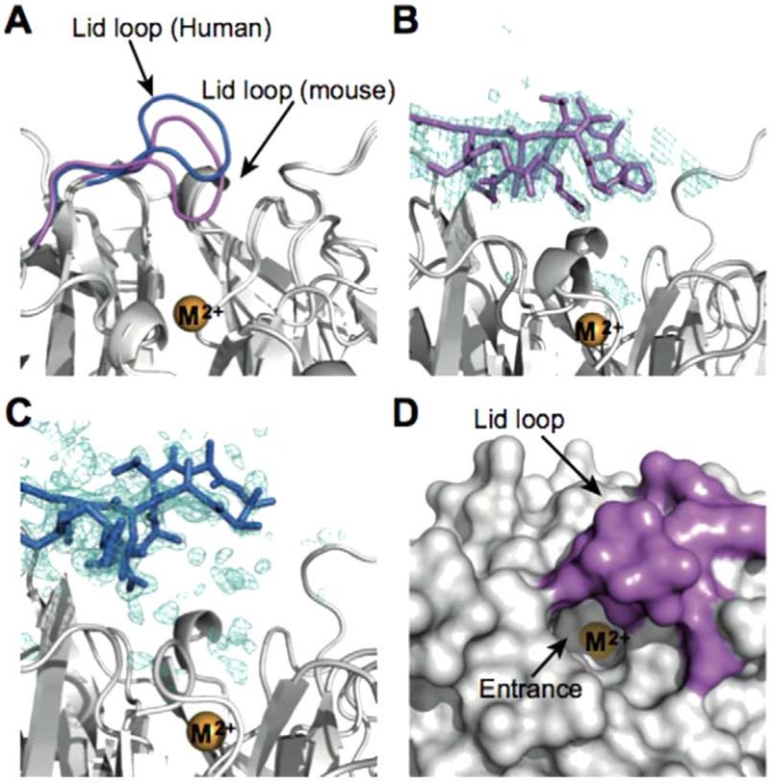


Figure 23 - A) Lid loops of mouse and human SMP30/GNL in the substrate free form are shown in purple and blue, respectively. The divalent metal ion (labeled as M2+) is shown in orange. (B, C) SAomit maps (mFo-DFc maps) for the lid loop residues in mouse (B) and human (C) SMP30/GNL. The contour levels of the SA-omit maps are 3.0 s and 2.0 s for panels B and C, respectively. (D) Surface representation of mouse SMP30/GNL around the lid loop. The entrance for the substrate-binding cavity is indicated by an arrow. Residues in the lid loop are shown in purple. Presented figure and description were both adapted from Aizawa *et al.* (2013).

3.1. *Regucalcin* loss in Nematodes

The reports of Misawa and Yamaguchi, (2000), strongly suggest that *Regucalcin* performs essential biological functions and later, Scott and Bahnson, (2011) concluded that *Regucalcin* is an essential gene, that assumes an important biological role in the calcium homeostasis and other important biological features in vertebrates. Nevertheless, the *Regucalcin* gene could not be found in any of the 35 analysed Nematode species. This is rather interesting, since calcium homeostasis is essential for many biological processes. In order to understand whether Nematodes regulate their calcium levels by a different mechamism, the model organism, *Caenorhabditis elegans*, was chosen for comparative interactome analysis. The first results showed that there were, in fact orthologues of the calcium homeostasis related interactors found in *D. melanogaster* (table 9) in *C. elegans*.

Table 9 – Calcium homeostasis related protein orthologues in *C. elegans*

Protein (*D. melanogaster*)	***C. elegans* GeneID (Orthologue)**	**Uniprot ID *C. elegans***	***C. elegans* Symbol**	**Size**
Chloride intracellular channel	173314	Q8WQA4	exc-4	290
FK506-binding protein 14	191634	Q20107	fkb-1	139
Supercoiling factor	180769	G5EBH7	calu-1	314
Calreticulin	178997	P27798	crt-1	395
Translationally controlled tumor protein	172944	Q93573	tct-1	181

Using firstly, four of these five orthologues, individual interactomes were made (191634, 180769, 178997 and 172944). After obtaining all its interactors, these four lists were intercepted. This intersection had the objective of finding common interactors to all of these genes, in a way that it was possible to find gene candidates that may be assuming the role of *Regucalcin* in Nematodes. Nevertheless, these preliminary results identified four putative candidates that can be seen represented in table 10.

Table 10 – Results of the intersection of the calcium homeostasis related orthologues interactomes in *C. elegans* (191634, 180769, 178997, 172944).

Interactor Gene ID	**Uniprot ID**	**Name**	**Function**	**Size**
171646	H2L0Q3	Gamma-aminobutyric acid type B receptor subunit 1	Component of a heterodimeric G-protein coupled receptor for GABA, formed by gbb-1 and gbb-2 (By similarity). Within the heterodimeric GABA receptor, only gbb-1 seems to bind agonists, while gbb-2 mediates coupling to G proteins (By similarity). Ligand binding causes a conformation change that triggers signaling via guanine nucleotide-binding proteins (G proteins) and modulates the activity of down-stream effectors, such as adenylate cyclase (By similarity). Signaling inhibits adenylate cyclase, stimulates phospholipase A2, activates potassium channels, inactivates voltage-dependent calcium-channels and modulates inositol phospholipid hydrolysis (By similarity). Calcium is required for high affinity binding to GABA (By similarity). Plays a critical role in the fine-tuning of inhibitory synaptic transmission (By similarity). Pre-synaptic GABA receptor inhibits neurotransmitter release by down-regulating high-voltage activated calcium channels, whereas postsynaptic GABA receptor decreases neuronal excitability by activating a prominent inwardly rectifying potassium (Kir)	899

			conductance that underlies the late inhibitory postsynaptic potentials (By similarity). Along with gbb-2, may couple to the G(o) alpha G-protein goa-1 to negatively regulate cholinergic receptor activity in the presence of high levels of acetylcholine in ventral cord motor neurons. As acetylcholine depolarizes body wall muscles, modulation of acetylcholine levels most likely results in the control of locomotory behavior. Acts in neurons to regulate lifespan, and this may be through G-protein-egl-8/PLC-beta signaling to the transcription factor daf-16/FOXO.	
178944	O44158	Related to yeast Vacuolar Protein Sorting factor	Endosome transport via multivesicular body sorting pathway. Protein transport. Receptor catabolic process.	234
178989	O16369	AP complex subunit sigma	Apical protein localization. intracellular protein transport. vesicle-mediated transport.	157
179598	Q22836	Uncharacterized protein T27F2.1	Embryo development ending in birth or egg hatching. germ cell development. locomotion. molting cycle. mRNA splicing, via spliceosome. nematode larval development. oviposition. positive regulation of transcription, DNA-templated. regulation of gene expression. uterus and vulval development.	535

After this first interception was made another separate orthologue gene combination was intercepted, this time using the genes 178614, 191634, 178997 and 172944. These were once again calcium homeostasis related orthologues interactomes in *C. elegans.* In addition, like in the previous case our goal was to establish an interaction network that could lead to the finding of a candidate gene that could be assuming the *Regucalcin*'s function. The results are shown in the table 11.

Table 11 - Results of the intersection of the calcium homeostasis related orthologues interactomes in *C. elegans* (178614, 191634, 178997 and 172944).

Interactor Gene ID	Uniprot ID	Name	Function	Size
177856	Q18680	Inorganic pyrophosphatase 1, EC 3.6.1.1 (Pyrophosphate phospho-hydrolase, PPase)	Catalyzes the hydrolysis of inorganic pyrophosphate (PPi) forming two phosphate ions. Plays a role in intestinal development and subsequent normal secretory, digestive and absorption functions. Required for larval development.	427
180028	P52011	Peptidyl-prolyl cis-trans isomerase 3, PPIase 3, EC 5.2.1.8 (Cyclophilin-3) (Rotamase 3)	PPIases accelerate the folding of proteins. It catalyzes the cis-trans isomerization of proline imidic peptide bonds in oligopeptides.	173

After the results of the previous interceptions (represented in tables 10 and 11), it was found that there were in fact interactors of those intercepted genes that were once again involved with calcium homeostasis.

As it was not possible to find an evident Regucalcin replacement gene candidate, for the interactors 171646, 178944, 178989 and 179598, of *C. elegans*, their homologous genes were found in *D. melanogaster* being, respectively 36409, 39881, 42835 and 31840, to further investigate if the calcium interaction networks were shared in these two protostomes.

For these fly homologous genes an individual interactome was also made, and only the second level calcium interactors were transcribed to table 12.

Table 12 – second term calcium interactors of *C. elegans* in *D. melanogaster*

Gene ID	Uniprot ID	Name	Function	Size
33019	P22465	Annexin B10	Calcium-dependent phospholipid binding, calcium ion binding.	321
34170	Q9VLL3	A-kinase anchor protein 200	Scaffolding protein involved in the regulation of PKA signaling and anchoring to the actin cytoskeleton integrating signals propagated by cAMP, diacylglycerol and calcium. Contributes to	753

				the maintenance and regulation of cytoskeletal structures in germline via PKA-mediated signaling. As part of ethanol response in the glia, mediates ethanol-induced structural remodeling of actin cytoskeleton and perineurial membrane topology by anchoring PKA to the membrane of perineurial glia. In specific tissues such as eye and thorax, promotes N/Notch protein stability by inhibiting Cbl-mediated ubiquitination and lysosomal degradation pathway of N/Notch in a PKA-independent way. In the circadian brain neurons evening cells (E-cells), might have a role in circadian pacemaker synchronization by playing a redundant role in signaling downstream of the G protein-couple receptor Pdfr.	
Interactors IDs of (37922)					
	34363	A8DY93	Slowpoke 2, isoform D	Calcium-activated potassium channel activity, intracellular sodium activated potassium channel activity, outward rectifier potassium channel activity.	1,878
	34193	P33438	Glutactin	Binds calcium ions. defense response to other organism, motor neuron axon guidance, synaptic target inhibition.	1,026
	37111	A1ZB71	Glutathione S transferase E6	Glutathione transferase activity.	222
	37449	A0A0B4K7C5	Peptidylprolyl isomerase, EC 5.2.1.8	Calcium ion binding, peptidyl-prolyl cis-trans isomerase activity.	220
	40451	P19889	60S acidic ribosomal protein P0	Calcium ion binding Source: CAFA class I DNA-(apurinic or apyrimidinic site) endonuclease activity Source: UniProtKB-EC. class II DNA-(apurinic or apyrimidinic site) endonuclease activity. DNA-(apurinic or apyrimidinic site) endonuclease activity, endonuclease activity, magnesium ion binding, structural constituent of ribosome.	317
	53446	P06742	Myosin light chain alkali	Calcium ion binding, microfilament motor activity.	155

	41166	P29413	Calreticulin	Molecular calcium-binding chaperone promoting folding, oligomeric assembly and quality control in the ER via the calreticulin/calnexin cycle. This lectin may interact transiently with almost all of the monoglucosylated glycoproteins that are synthesized in the ER.	406
	40090	Q9VVX3	Frizzled-2, dFz2	Receptor for Wnt proteins. Most of frizzled receptors are coupled to the beta-catenin canonical signaling pathway, which leads to the activation of disheveled proteins, inhibition of GSK-3 kinase, nuclear accumulation of beta-catenin and activation of Wnt target genes. A second signaling pathway involving PKC and calcium fluxes has been seen for some family members, but it is not yet clear if it represents a distinct pathway or if it can be integrated in the canonical pathway, as PKC seems to be required for Wnt-mediated inactivation of GSK-3 kinase. Both pathways seem to involve interactions with G-proteins. Required to coordinate the cytoskeletons of epidermal cells to produce a parallel array of cuticular hairs and bristles.	694
	38413	Q8IRD3	Glutathione peroxidase	Glutathione peroxidase activity, peroxidase activity.	238
	38145	Q9W0H8	IP16409p (RH25118p) (Supercoiling factor, isoform A)	Calcium ion binding.	329
	35378	M9ND00	Dynamin associated protein 160, isoform D	Calcium ion binding, nucleic acid binding.	1,190
	33196	M9NER8	Dumpy, isoform I	Calcium ion binding, DNA-binding transcription factor activity, extracellular matrix structural constituent, zinc ion binding.	15,638
Interactors IDs of (39881)	32746	Q00963	Spectrin beta chain	Spectrin is the major constituent of the cytoskeletal network underlying the	2,291

				erythrocyte plasma membrane. It associates with band 4.1 and actin to form the cytoskeletal superstructure of the erythrocyte plasma membrane. Interacts with calmodulin in a calcium-dependent manner.	
	42314	Q59DP8	Calcium-transporting ATPase, EC 7.2.2.10	This magnesium-dependent enzyme catalyses the hydrolysis of ATP coupled with the transport of calcium.	1,120
Interactors IDs of (31840)	32168	Q59E33	LD21442p (SR-related CTD associated factor 6, isoform A)	RNA binding, cellular calcium ion homeostasis.	960
	37449	A0A0B4K7C5	Peptidylprolyl isomerase	Calcium ion binding.	220
Interactors IDs of (37922)	38362	Q9VZW1	Phospholipid scramblase	May mediate accelerated ATP-independent bidirectional transbilayer migration of phospholipids upon binding calcium ions that results in a loss of phospholipid asymmetry in the plasma membrane.	263
	44307	P48456	Serine/threonine-protein phosphatase 2B catalytic subunit 1	Calcium-dependent, calmodulin-stimulated protein phosphatase. This subunit may have a role in the calmodulin activation of calcineurin.	622

These results indicate that calcium homeostasis is still being regulated in Nematodes, due to the presence of the identified orthologues, likely using a biological mechanism that relies on an alternative protein other than Regucalcin. Nevertheless, further work is needed to describe this process in Nematodes.

4. Positively selected amino acid sites

The identification of a positively selected amino acid sites (PSS) can give insight into important features of *Regucalcin* and in phylogenetic groups where this gene is duplicated, it can provide useful information of the evolution on new functions.

Several groups showed signs of PSS, namely, in Aves (gene 1 and 2) (figure 24 and 25, respectively), in Formicoidea (gene 1, presented in figure 26), Sophophora – *Regucalcin* gene (figure

27) and Sophophora – *Dca* gene (figure 28), Apoidea (gene 2, in figure 29), Cyprinidae (gene C1", in figure 30), in the group of fish not containing the Cyprinidae and Salmonidae groups (figure 31). Lepidoptera (gene 2, figure 32), Reptilia (gene 1 and 2, in figures 33 and 34, respectively) and lastly, in the Mammalia group (gene 2, figure 35).

In what concerns gene 1 of Aves, it is possible to observe one amino acid that was positively selected when using 52 bird species. Thus, one representative sequence of this group was elected, namely that from *Coturnix japonica* (Japanese quail) belonging to the Phasianidae group (XP_015706795.1), and after inferring its 3D structure this site was mapped on its protein structure (figure 24).

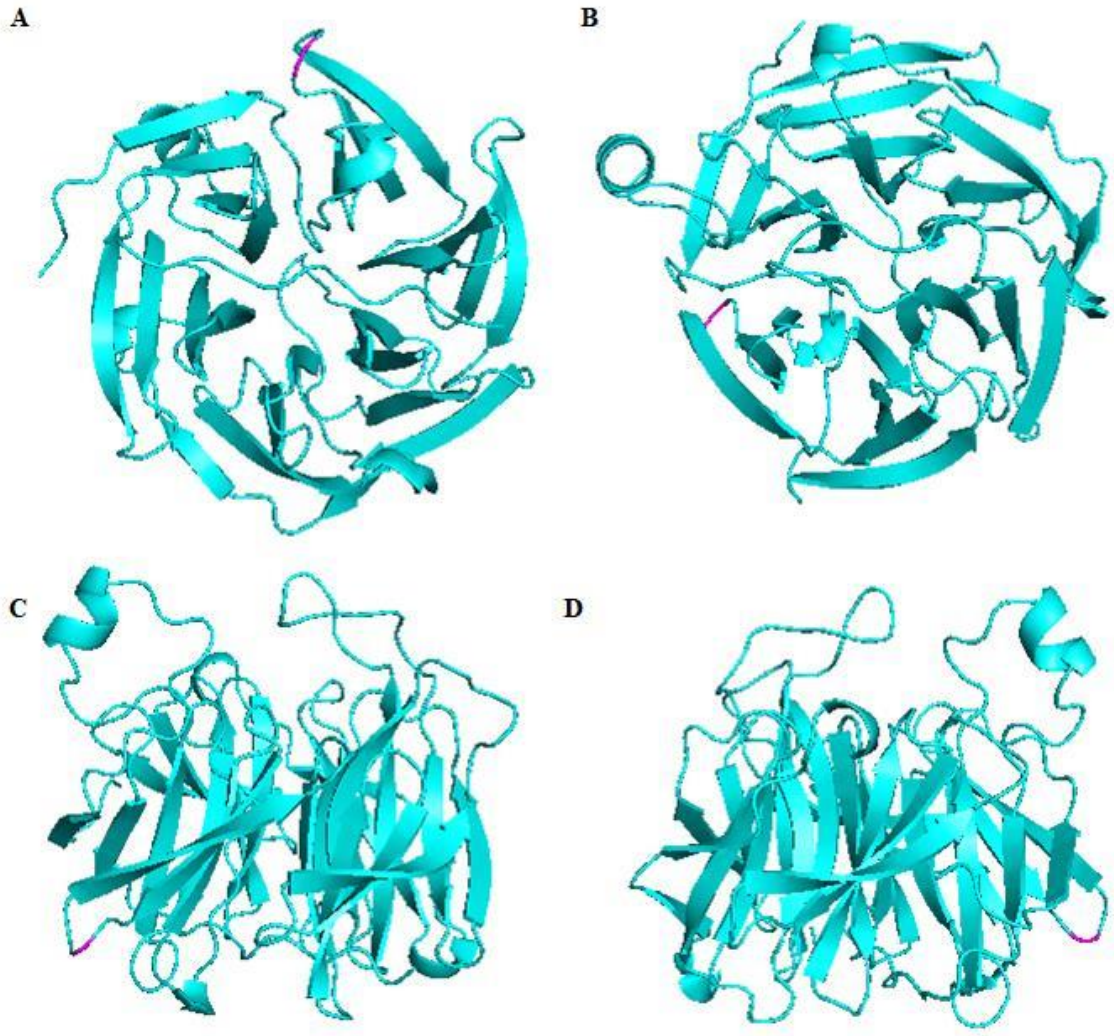


Figure 24 - Positively selected amino acid sites (in purple) in the Regucalcin protein structure of *Coturnix Japonica*.

Interestingly, it is possible to see that the positively selected amino acid is located on the newly identified protein interaction domain.

For gene 2 of Aves, there are eleven amino acids that were positively selected when using 48 bird species. Thus, one representative sequence of this group was elected, namely, *Gallus gallus* (domestic chicken) belonging to the Phasianidae group (XP_015133172.1), and after inferring its structure these sites were mapped on top of it (figure 25).

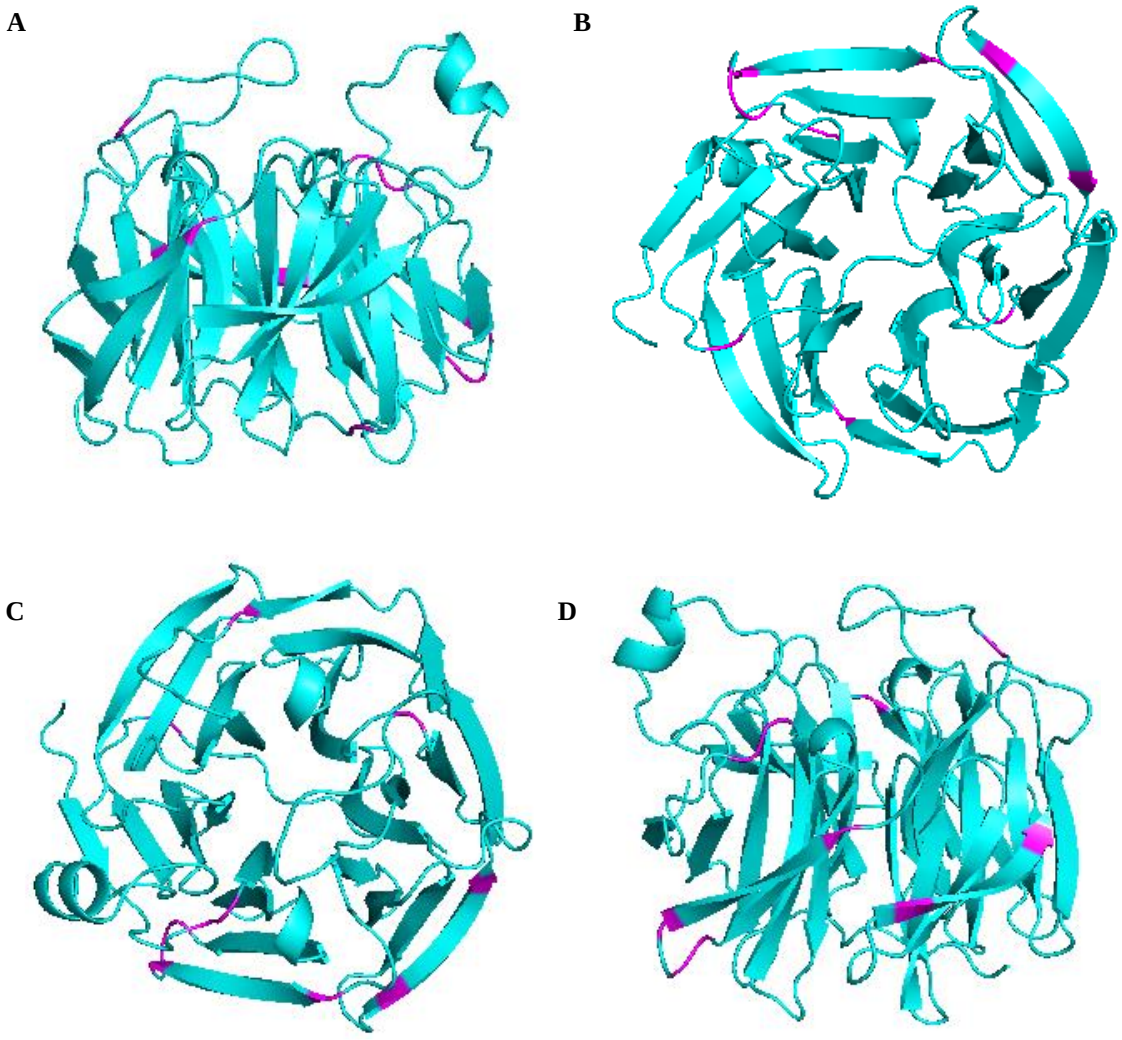


Figure 25 - Positively selected amino acid sites (in purple) in the Regucalcin protein structure of *Gallus gallus*

The PSS are located close to the Regucalcin protein lid but also in the newly identified protein interaction domain.

For gene 1 of Formicoidea, it is possible to observe five amino acids that were positively selected when using 15 species. Thus, one representative sequence of this group was elected, namely, *Acromyrmex echinatior* (Panamanian leafcutter ant) belonging to Formicidae (XP_011068254.1) and after inferring its 3D structure the sites were mapped on top of it (figure 26).

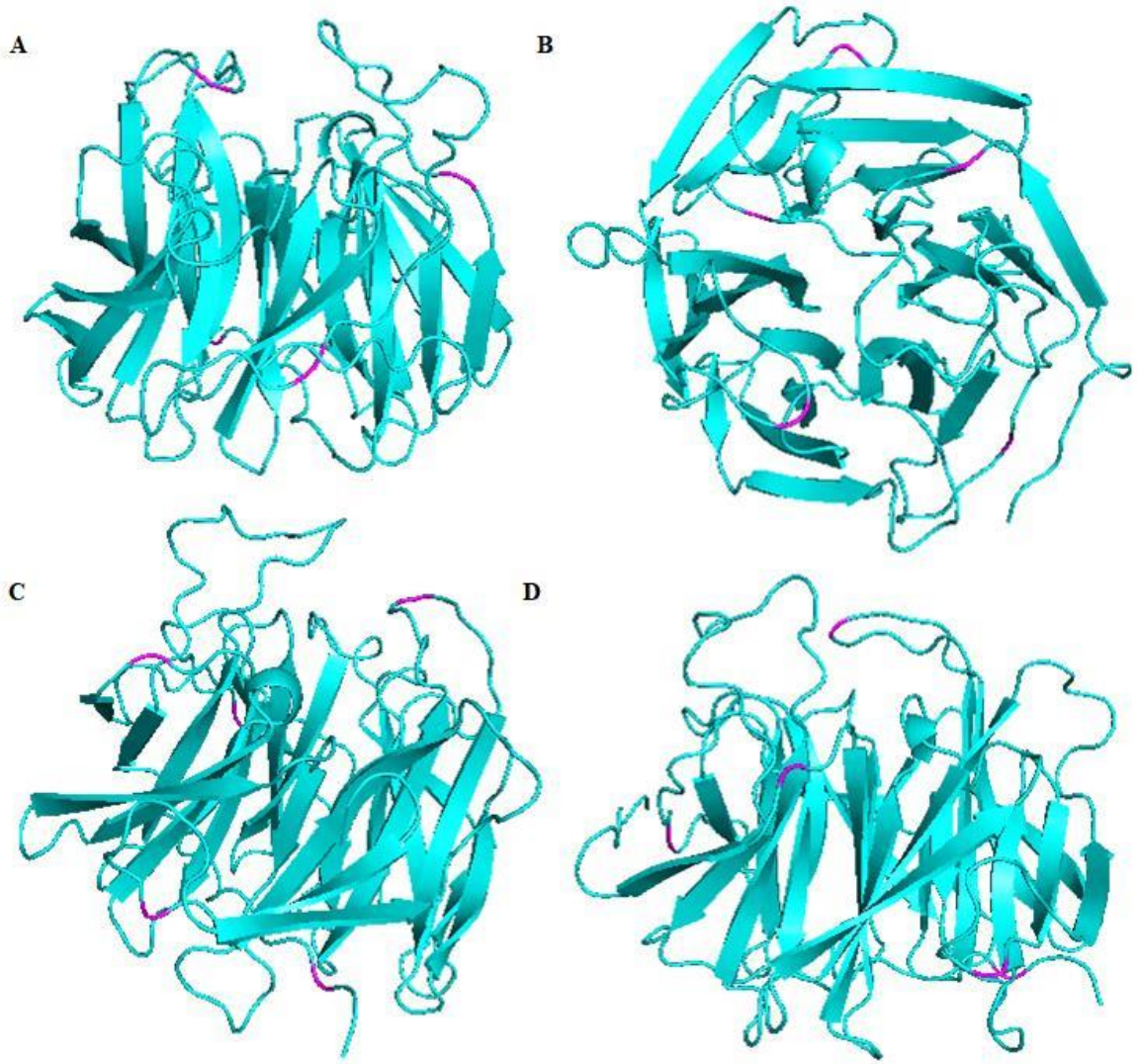


Figure 26 - Positively selected amino acid sites (in purple) in the Regucalcin protein structure of *Acromyrmex echinatior*.

Some PSS are present in the lid of the Regucalcin protein, and the others are around the newly proposed protein interaction domain.

For Sophophora-*Regucalcin* gene, it is possible to observe two amino acid sites that were positively selected when using 18 species. Thus, one representative sequence of this group was elected, namely, *Drosophila melanogaster* (fruit fly) belonging to Drosophilidae (AAN09306.2) and after inferring its 3D structure the sites were mapped on top of it (figure 27).

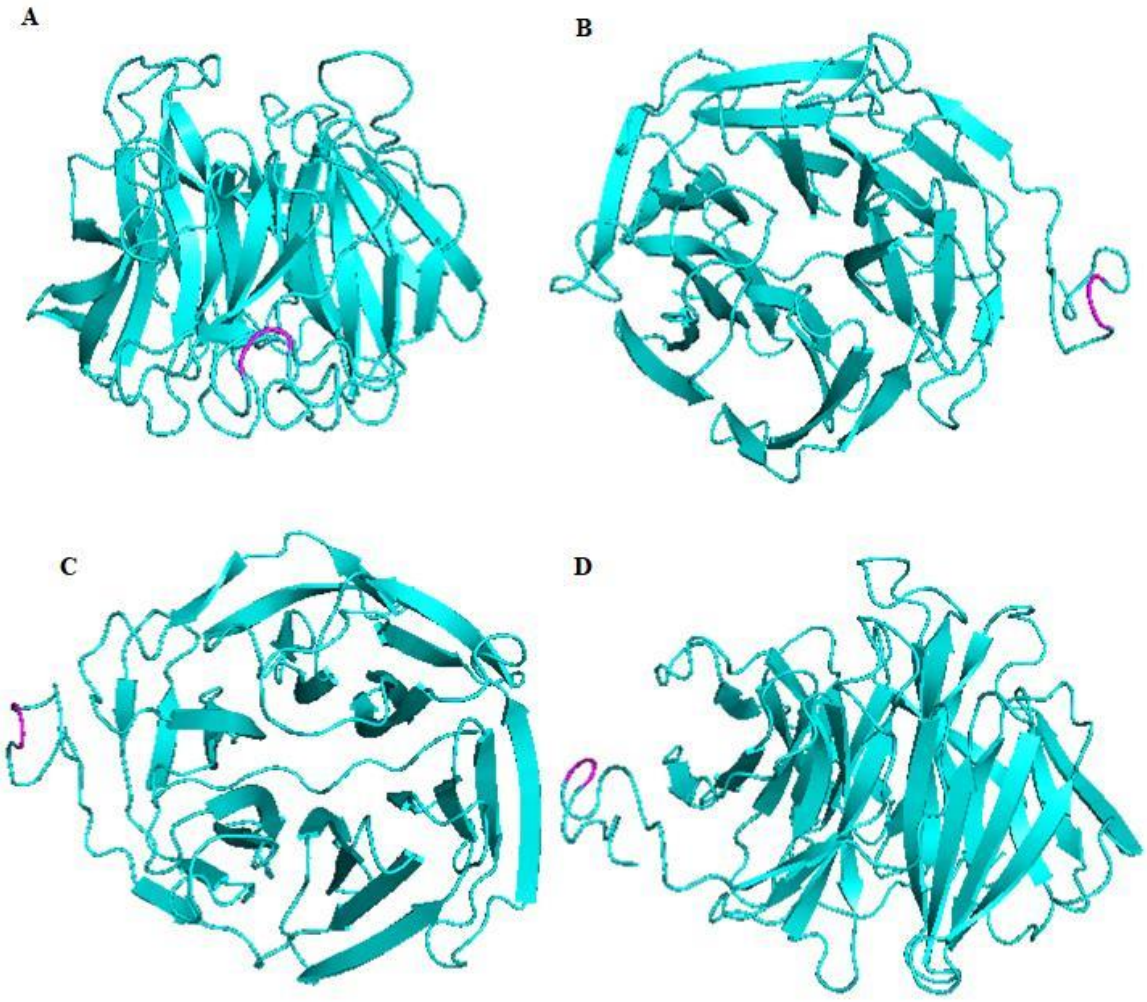


Figure 27 - Positively selected amino acid sites (in purple) in the Regucalcin protein structure of *Drosophila melanogaster*, *from Drosophila – Regucalcin group.*

The two positively selected amino acid sites are located close together on a lateral side of the molecule.

For Sophophora-*Dca* gene, it is possible to observe five amino acid sites that were positively selected when using 18 species. Thus, one representative sequence of this group was elected, namely, *Drosophila melanogaster* (fruit fly) belonging to Drosophilidae (AGB95961.1) and after inferring its 3D structure the sites were mapped on top of it (figure 28).

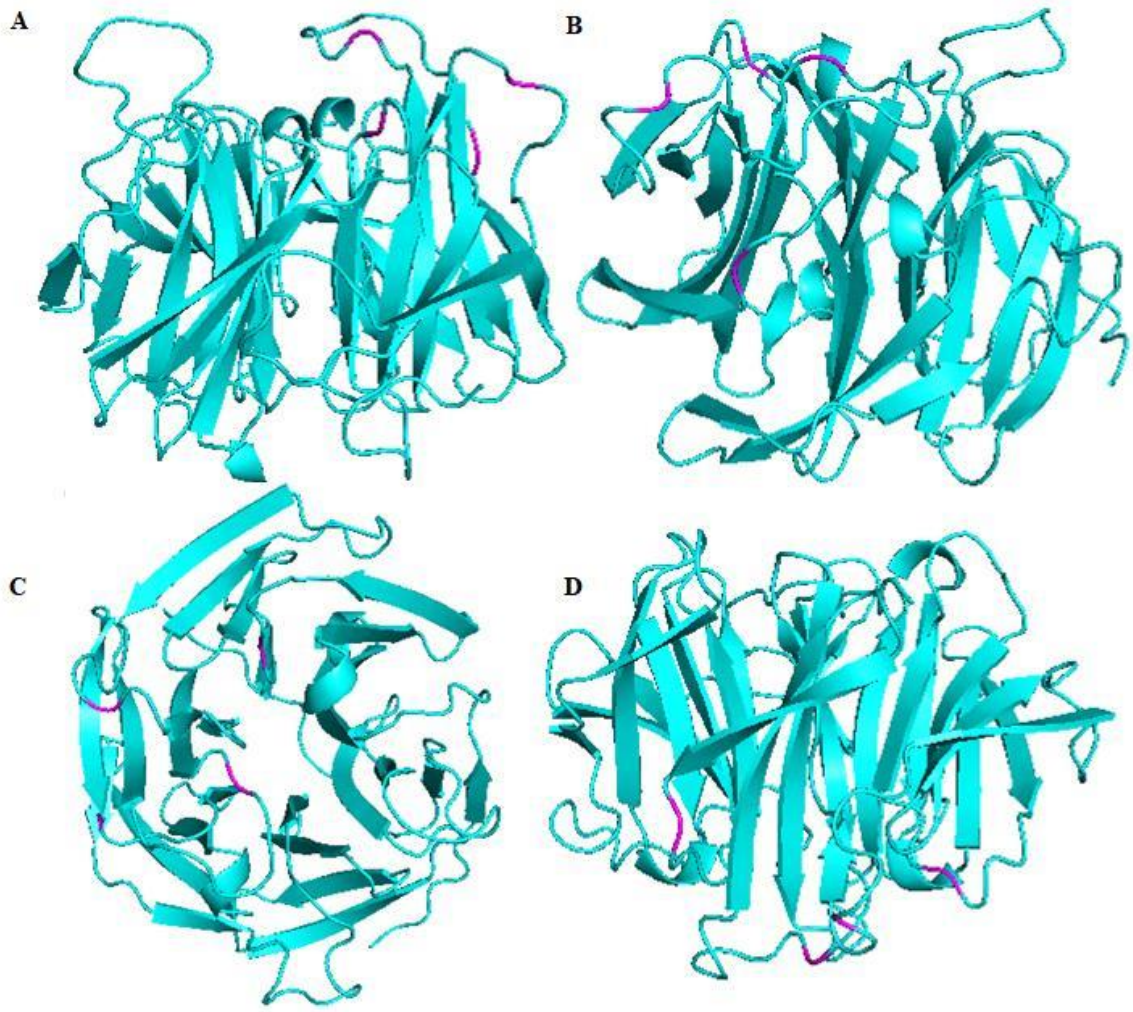


Figure 28 - Positively selected amino acid sites (in purple) in the Regucalcin protein structure of *Drosophila melanogaster, from Drosophila – Dca group.*

The five selected amino acid sites are close together around the newly identified putative protein interaction domain site and the protein's lid.

For gene 2 of Apoidea, it is possible to observe three amino acid sites that were positively selected when using 8 species. Thus, one representative sequence of this group was elected, namely, *Bombus terrestris* (buff-tailed bumblebee) belonging to Apidae (XP_020722941.1) and after inferring its 3D structure the sites were mapped on top of it (figure 29).

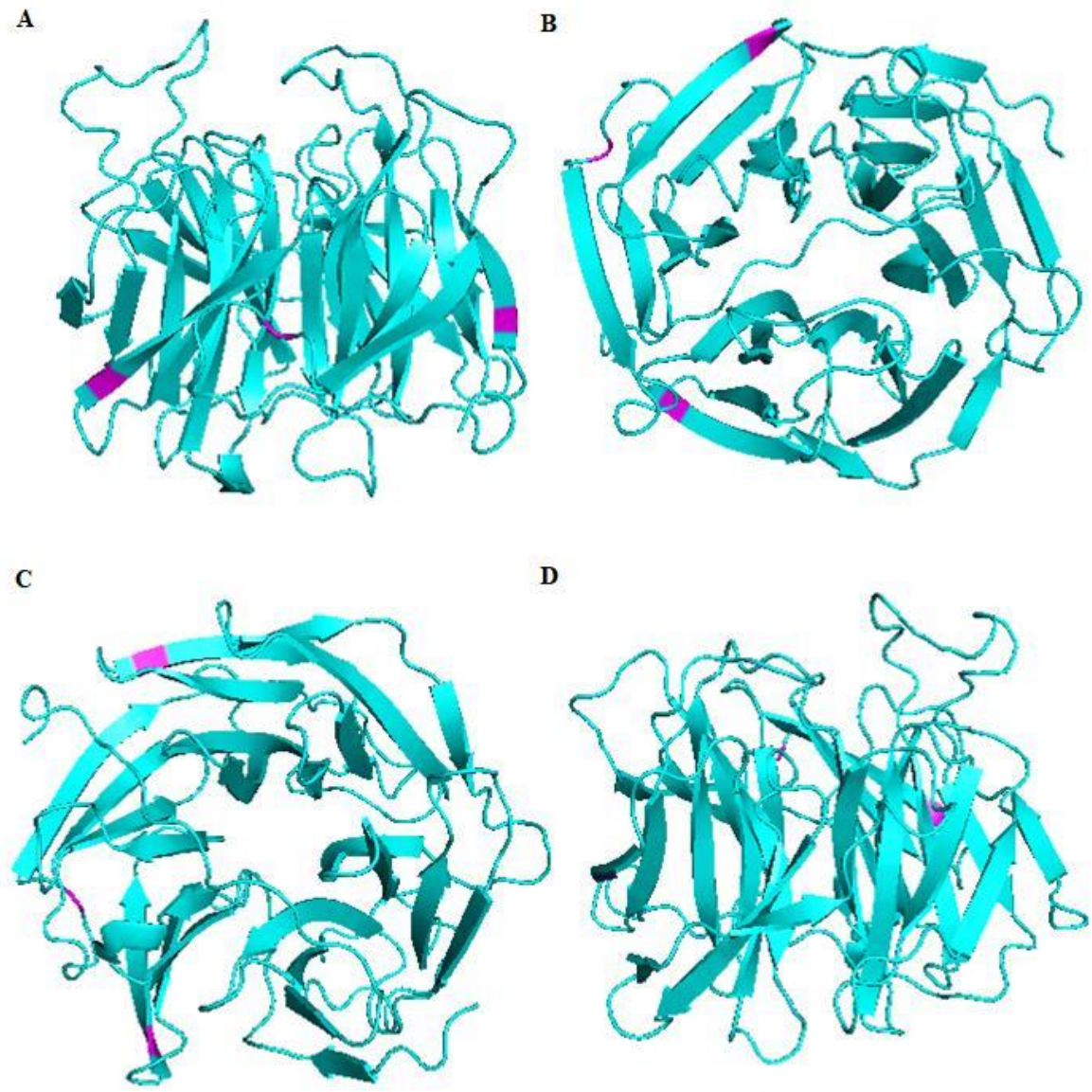


Figure 29 - Positively selected amino acid sites (in purple) in the Regucalcin protein structure of *Bombus terrestris.*

The three selected amino are placed in the same molecule's side, in the newly protein interaction domain.

For Cyprinidae gene C1", it is possible to observe one amino acid site that was positively selected when using 5 species. Thus, one representative sequence of this group was elected, namely, *Danio rerio* (Zebra fish) belonging to Cyprinidae (NP_991309.1) and after inferring its 3D structure the site was mapped on top of it (figure 30).

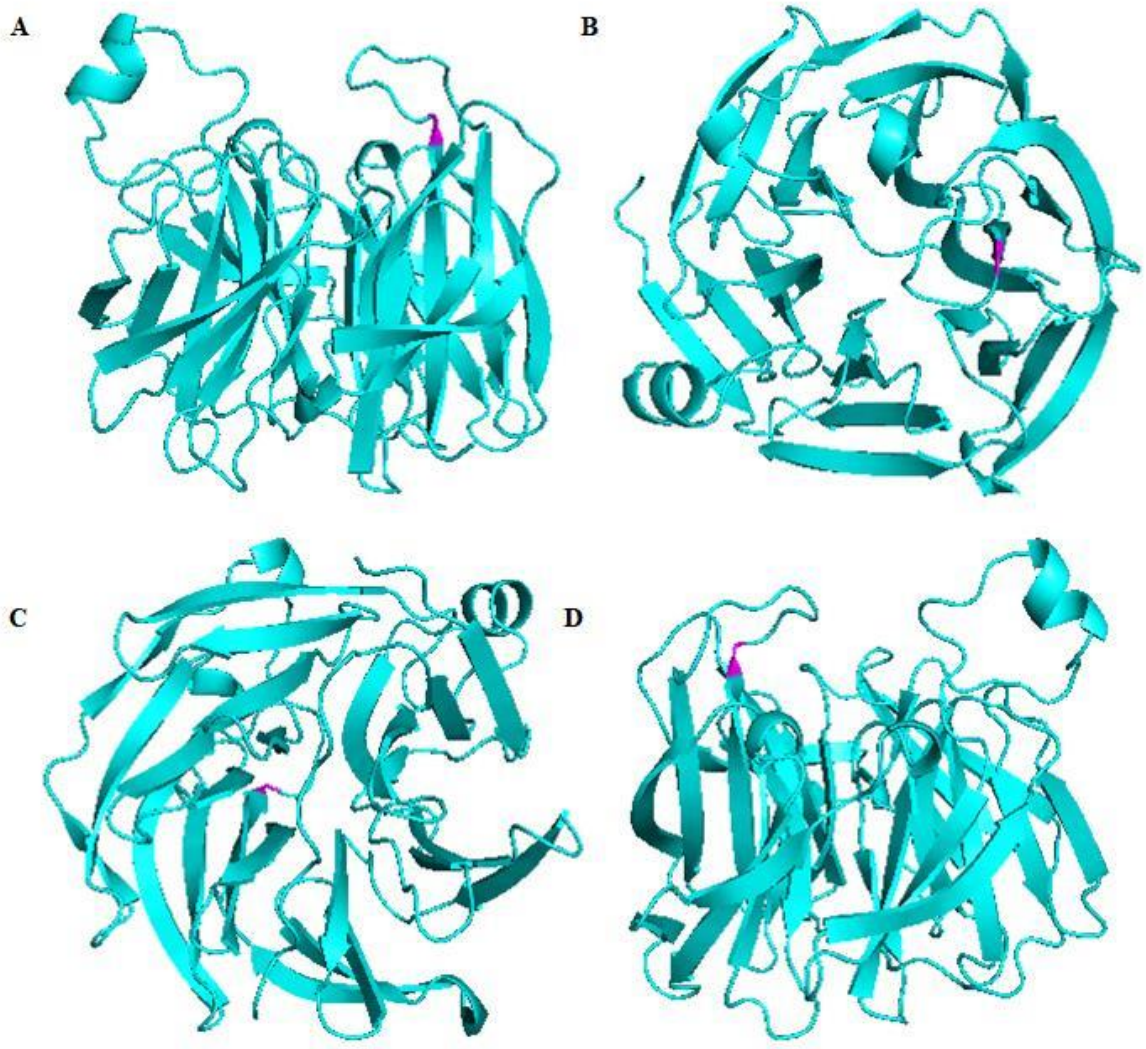


Figure 30 - Positively selected amino acid ste (in purple) in the Regucalcin protein structure of *Danio rerio.*

It is possible to observe that the selected amino acid site it is placed in the protein's lid.

For the case of Fish without Cyprinidae and Salmonidae, it is possible to observe one amino acid site that was positively selected when using 30 species. Thus, one representative sequence of this group was elected, namely, *Poecilia latipinna* (sailfin molly) belonging to Poeciliidae (XP 014904780.1) and after inferring its 3D structure the site was mapped on top of it (figure 31).

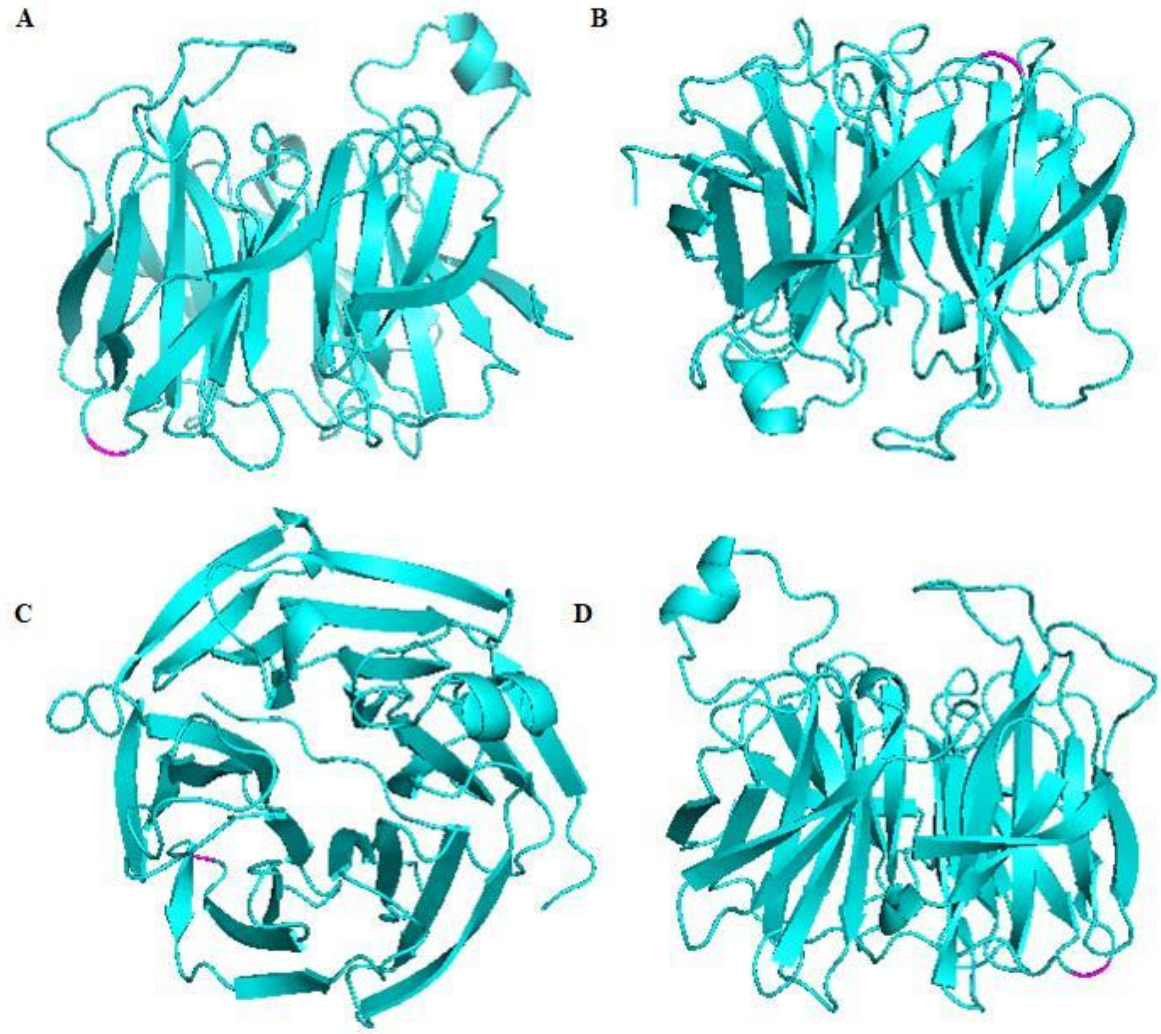


Figure 31 - Positively selected amino acid site (in purple) in the Regucalcin protein structure of *Poecilia latipinna*.

The positively selected amino acid site is placed in the opposite side of protein's lid.

For Lepidoptera gene 2, it is possible to observe one amino acid site that was positively selected when using 8 species. Thus, one representative sequence of this group was elected, namely, *Bombyx mori* (domestic silkworm) belonging to Bombicidae (XP_012549788.1) and after inferring its 3D structure the site was mapped on top of it (figure 32).

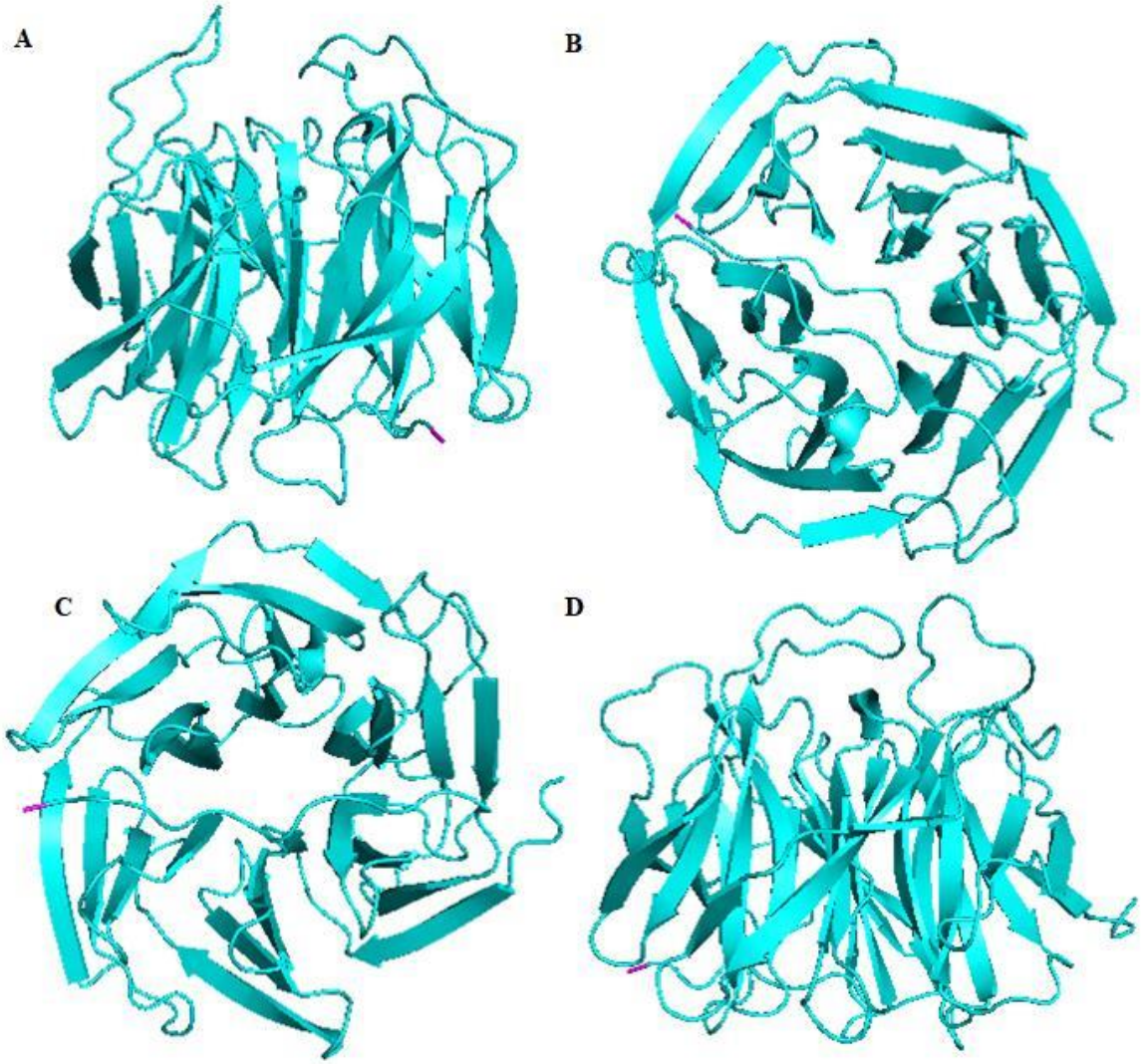


Figure 32 - Positively selected amino acids (in purple) in the Regucalcin protein structure of *Bombyx mori*

The selected amino acid site is placed in a lateral side in the opposite direction of the protein's lid.

For Reptilia gene 1, it is possible to observe one amino acid site that was positively selected when using 8 species. Thus, one representative species of this group was elected, namely, *Gekko japonicus* belonging to Gekkonidae (XP_015266456.1) and after inferring its 3D structure the site was mapped on top of it (figure 33).

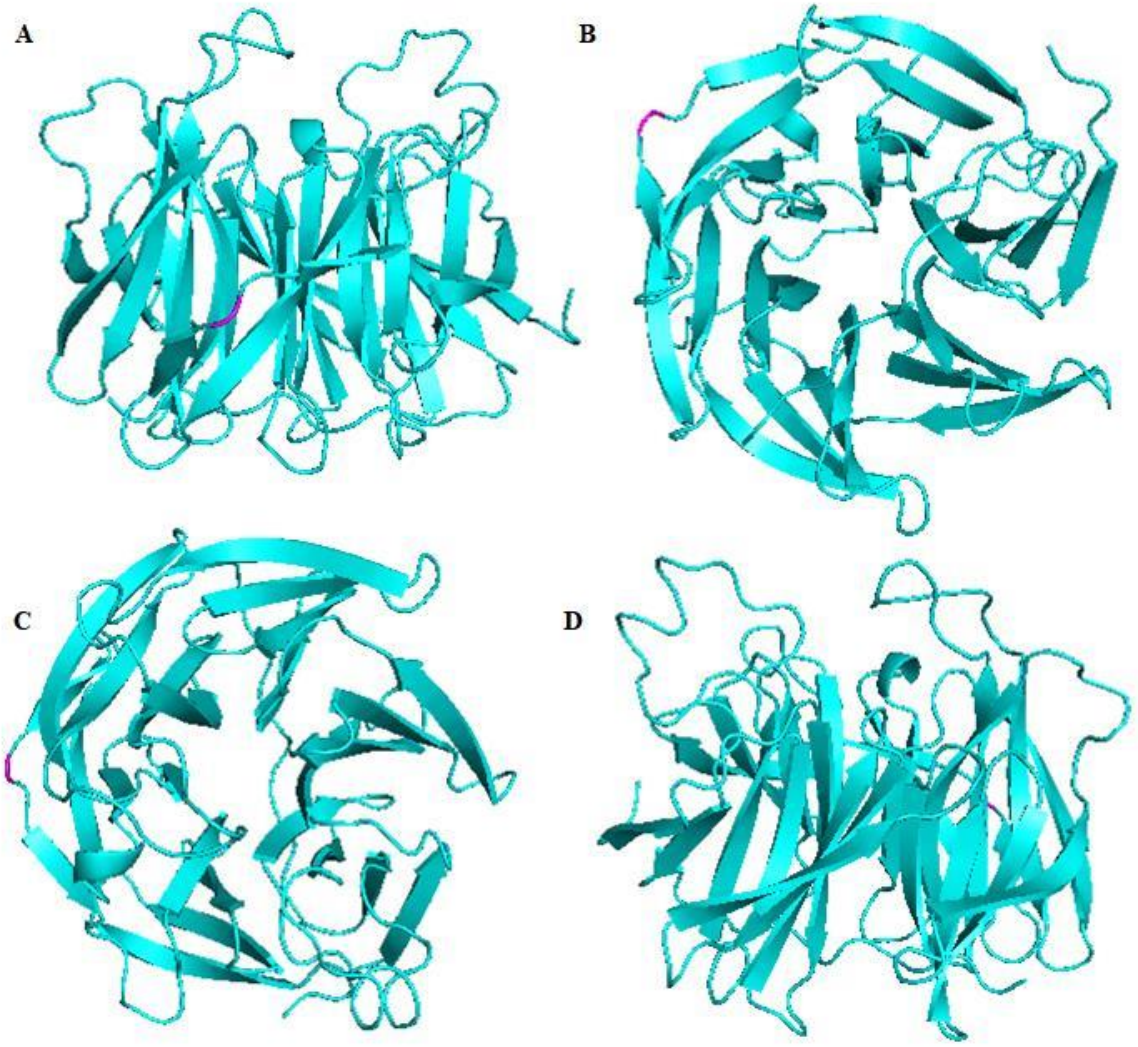


Figure 33 - Positively selected amino acid site (in purple) in the Regucalcin protein structure of *Gekko japonicus*

The selected amino acid site is placed in the lateral side of the molecule, in the inferred protein interaction domain.

For Reptilia gene 2, it is possible to observe one amino acid site that was positively selected when using 10 species. Thus, one representative sequence of this group was elected, namely, *Gekko japonicus* belonging to Gekkonidae (XP_015266450.1) and after inferring its 3D structure the site was mapped on top of it (figure 34).

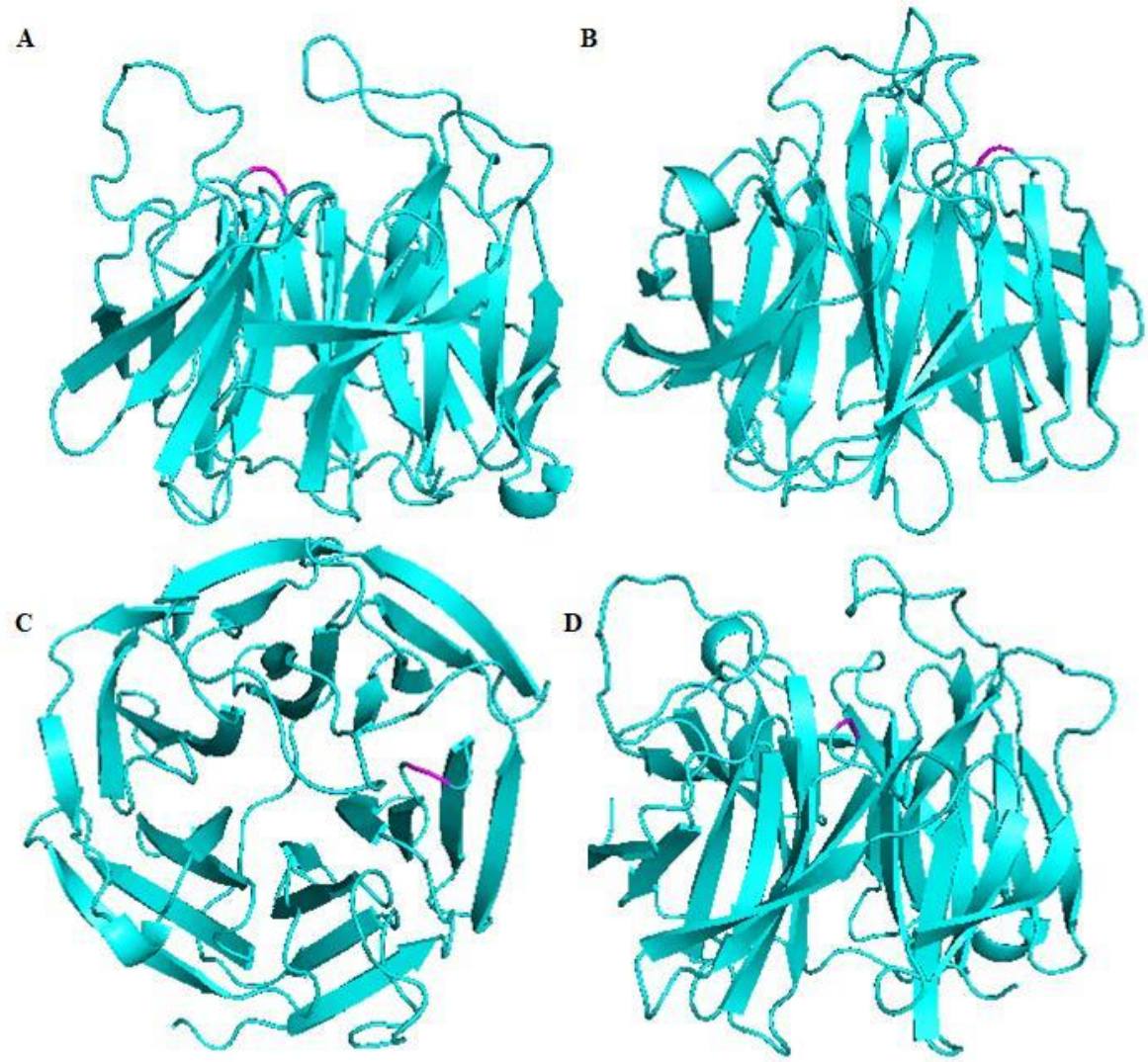


Figure 34 - Positively selected amino acid site (in purple) in the Regucalcin protein structure of *Gekko japonicas.*

The selected amino acid site is placed in a central position really close to the protein's lid.

For Mammalia gene 2, six amino acids sites were inferred by the FUBAR analysis, yet only three can be seen here for this species (due to alignment gaps) that were positively selected when using 100 species. Thus, one representative sequence of this group was elected, namely, *Homo sapiens* (human) belonging to Hominidae (NP_690608.1) and after inferring its 3D structure the sites were mapped on top of it (figure 35).

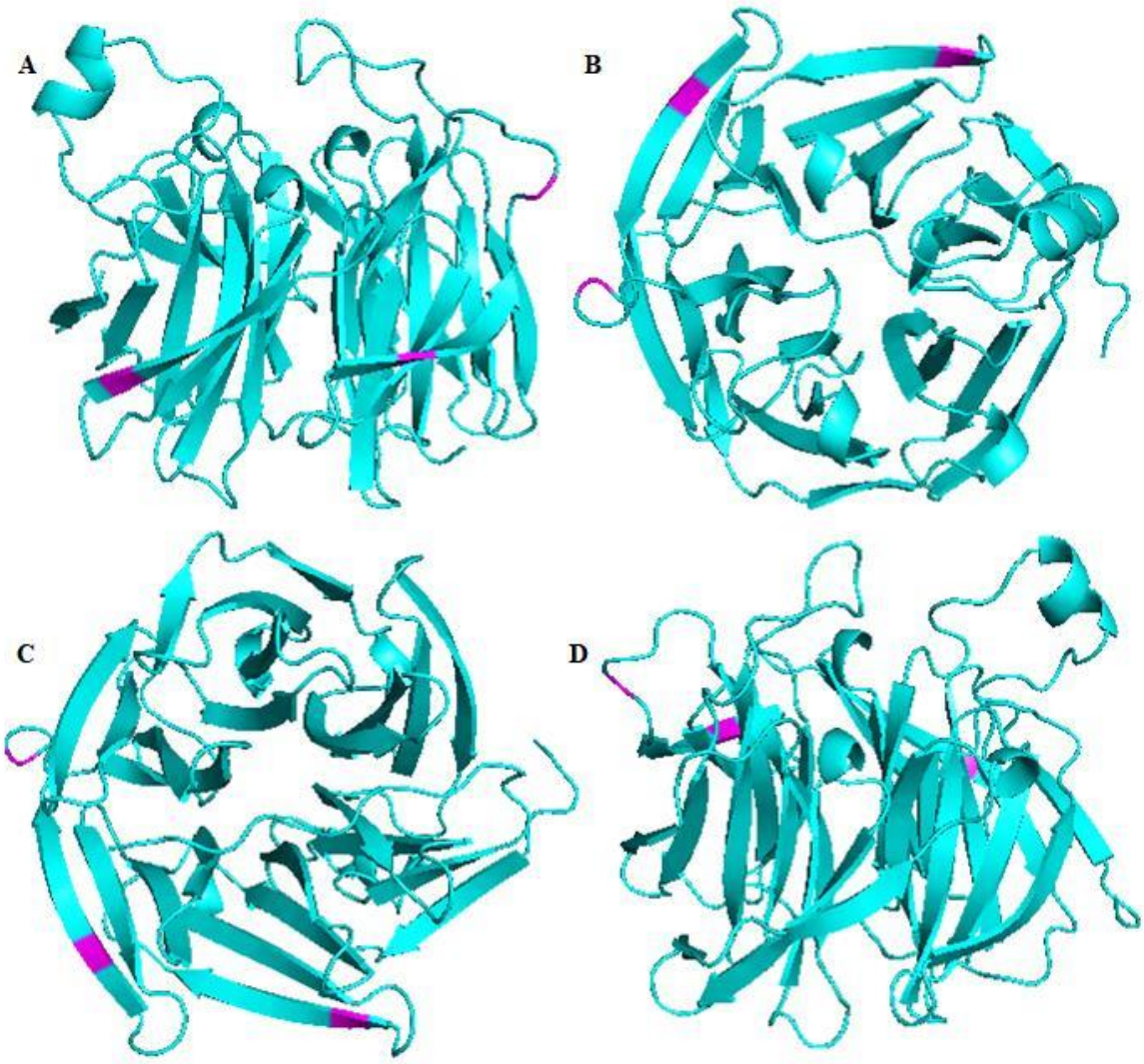


Figure 35 - Positively selected amino acid sites (in purple) in the Regucalcin protein structure of *Homo sapiens.*

The selected amino acids are placed close to the protein's lid as well as in a lateral position of the molecule.

Overall, looking to the 12 analysed groups it is possible to observe that in seven of them, there are both positively selected amino acid sites in the protein's lid and in the newly inferred interaction domain (having four of these groups both of the mentioned places with PSS). Only the groups Sophophora – *Regucalcin* gene and Fish without Cyprinidae and Salmonidae are missing both of these locations. Therefore, most PSS are predicted to alter either the catalytic activity of Regucalcin or regulate its binding activity with other proteins, and thus be involved in the emergence of new functions.

Discussion:

Phylogenetic analyses:

One of the main findings of this thesis is that the *Regucalcin* gene is not present in all animals and that it was lost multiple times independently. These were really surprising facts, due to the relevance of this gene being unquestionable by the power of its polyvalence of biological functions (for instance calcium homeostasis, AA biosynthesis and cell apoptosis).

It made us question how did they managed to evolve and regulate their calcium intake and Vitamin C synthesis, if they didn´t present this gene.

It seems likely that other genes replaced the functions played by the *Regucalcin* gene in these cases. Such genes are still unknown and could be an interesting study for the future.

The second aspect is that *Regucalcin* is often duplicated in Protostomes, but not in Deuterostomes. In evolution, gene duplication is the fastest way for adaptation.

Nevertheless, sometimes, when one of the gene duplicates experiences degenerate mutations, these end up reducing their joint levels and patterns of activity to the ones of the single ancestral gene (subfunctionalization). But, this would mean that in Protostomes subfunctionalization is more frequent. When Protostomia species are observed in detail, it is possible to see that the concentration of *Regucalcin* gene duplications mainly focuses on Insecta. Which by coincidence, is the invertebrate dominant population of the planet Earth (DeLong, 1990), meaning that they perhaps survived most of the mass extinctions, through adaptation until the present day and that these duplications might have contributed for it.

Lastly, the phylogenetic analysis shows that it is possible to establish evolutionary relationships between the presence/loss of the *Regucalcin*, *GULO* and *SVCT* genes. Two cases stood

out, one in the Protostomes and the other in the Deuterostomes. The first one had to do with the Nematoda group (represented by 35 species) in which *Regucalcin* and the *GULO* genes are absent, but three *SVCT*s were found. Nematodes may synthetize AA by an alternative pathway but this does not explain how did they perform the functions usually led by *Regucalcin*. As for the second case, the Monotremata group represented by one and unique living species the *Ornithorhynchus anatinus* were it had lost both of the *Regucalcin* gene duplicates and presented the *GULO* gene, yet no SVCTs. Therefore, it can be inferred that this species may produce its own AA yet does not has a way of transporting it in its organism, which is rather confusing. In addition to this, how does this organism performs the *Regucalcin* attributed functions if it lost the two gene duplicates, or did it create another calcium regulating alternative pathway, the question remains. This could be an interesting subject to address in future work.

Interaction domain:

After an intensive investigation of interactomes of calcium homeostasis related interactors, it was found that, in *D. melanogaster*, all the Regucalcin interacting proteins seem to interact in a specific place of the Regucalcin molecule. Therefore, a prediction of a putative protein interaction domain was made, out of this location.

On the other hand, even though the fact of Nematodes lacking *Regucalcin*, they share the same interactors (calcium homeostasis) with *D. melanogaster*, meaning that there is homology between the calcium homeostasis related interactors. This could also mean that Nematodes control indeed their calcium intake in a similar way to that of other protostomes containing *Regucalcin*, like *D. melanogaster*.

***Regucalcin* and *Dca* are essential genes:**

Both *Regucalcin* and *Dca* genes, in *Drosophila melanogaster*, are in fact essential genes. When looking to our fly experiments it was possible to observe that most of the pupae didn´t hatched and that the expected phenotypes were not found. This can mean that in opposition to what was once thought, both *RGN* and *Dca* are lethal in several developmental stages (egg and pupation), showing that they are in fact crucial for the biological mechanisms and life of this organism, and that *Dca* plays a more important role than that the literature describes. Therefore, *Dca* has to be more than just a mere cold acclimation gene, which also influences wing disks in *D. melanogaster*, and that there are many other functions to be revealed.

Positively selected amino acid sites:

After an extensive analysis to all the gene sequences of the analysed species, it was possible that in some groups, several amino acid sites were being positively selected. These groups were Aves gene 1 and 2, Formicoidea gene 1, Sophophora – *Regucalcin* gene, Sophophora – *Dca* gene, Apoidea gene 2, Cyprinidae gene C1", the group of fish not containing the Cyprinidae and Salmonidae groups, Lepidoptera gene 2, Reptilia gene 1 and 2 and lastly, in the Mammalia group gene 2.

A posterior observation of each of these selected amino acids in each of the 3D protein structures showed that there was a positive correlation between the proximity of these amino acids both with the putative interaction domain and with the protein lid. This indicates that the observed amino acid changes could indeed change the catalytic activity of Regucalcin as well as its ability to interact with its partners.

Conclusions:

Phylogenetic analyses:

Our phylogenetic analyses overall showed four important aspects. The first one is that the *Regucalcin* gene is not present in all animals. The second is that *Regucalcin* is often duplicated in Protostomes, but not in Deuterostomes. The third is that *Regucalcin* was lost multiple times independently and the last one is that it is possible to establish an evolutionary correlation between the presence/loss of the *Regucalcin/GULO/SVCT* genes in the animal kingdom.

***Regucalcin* and *Dca* are essential genes:**

It can be inferred that both *Regucalcin* and *Dca*, in the Sophophora subgenus are in fact essential genes, this because they are both lethal in several developmental stages (egg and pupation), meaning to be crucial for the biological mechanisms and life of this organism.

Interaction domain:

After an intensive investigation of interactomes of calcium homeostasis related interactors, it was found that, on one side in *D. melanogaster*, all the Regucalcin interacting proteins seem to interact in one specific place of the molecule, which can be suggestive of an interaction domain.

Regardless the fact that Nematodes lack *Regucalcin*, they share the same interactors with *D. melanogaster*, meaning that there is homology between the calcium homeostasis related interactors. This could also mean that Nematodes control indeed their calcium intake in a similar way other protostomes containing *Regucalcin* do.

Positively selected amino acid sites:

There are amino acids being selected over the evolution of the Regucalcin gene in the animal kingdom. These were overlapped on the Regucalcin protein structure and it was observed that they were mostly placed close to the protein's lid or to the newly identified interaction domain.

References:

Aizawa, S. *et al.* (2013) 'Structural Basis of the c-Lactone-Ring Formation in Ascorbic Acid Biosynthesis by the Senescence Marker Protein-30/Gluconolactonase', *PLOS ONE*, 8(1), p. 11.

Amano, A. *et al.* (2014) 'Effect of ascorbic acid deficiency on catecholamine synthesis in adrenal glands of SMP30/GNL knockout mice', *European Journal of Nutrition*, 53(1), pp. 177–185. doi: 10.1007/s00394-013-0515-9.

Arboleda-Bustos, C. E. and Segarra, C. (2011) 'The Dca Gene Involved in Cold Adaptation in Drosophila melanogaster Arose by Duplication of the Ancestral regucalcin Gene', *Molecular Biology and Evolution*, 28(8), pp. 2185–2195. doi: 10.1093/molbev/msr040.

Bürzle, M. *et al.* (2013) 'The sodium-dependent ascorbic acid transporter family SLC23', *Molecular Aspects of Medicine*, 34(2–3), pp. 436–454. doi: 10.1016/j.mam.2012.12.002.

Chen, S., Zhang, Y. E. and Long, M. (2010) 'New Genes in Drosophila Quickly Become Essential', *Science*, 330(6011), pp. 1682–1685. doi: 10.1126/science.1196380.

Dehal, P. and Boore, J. L. (2005) 'Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate', *PLoS Biology*. Edited by P. Holland, 3(10), p. e314. doi: 10.1371/journal.pbio.0030314.

DeLong, D. M. (1990) 'Man in a World of Insects', 60(4), p. 14.

Dietzl, G. *et al.* (2007) 'A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila', *Nature*, 448(7150), pp. 151–156. doi: 10.1038/nature05954.

Drouin, G., Godin, J.-R. and Page, B. (2011) 'The Genetics of Vitamin C Loss in Vertebrates', *Current Genomics*, 12(5), pp. 371–378. doi: 10.2174/138920211796429736.

Duque, P. M. B. (2018) 'Evolution of the vitamin C biosynthetic pathway and transport mechanism: from the global perspective to a Drosophila melanogaster case study'. Available at: https://repositorio-aberto.up.pt/handle/10216/118834 (Accessed: 1 August 2019).

Glasauer, S. M. K. and Neuhauss, S. C. F. (2014) 'Whole-genome duplication in teleost fishes and its evolutionary consequences', *Molecular Genetics and Genomics*, 289(6), pp. 1045–1060. doi: 10.1007/s00438-014-0889-2.

Gomi, K., Hirokawa, K. and Kajiyama, N. (2002) 'Molecular cloning and expression of the cDNAs encoding luciferin-regenerating enzyme from Luciola cruciata and Luciola lateralis', *Gene*, 294(1–2), pp. 157–166. doi: 10.1016/S0378-1119(02)00764-3.

He, X. and Zhang, J. (2005) ‘Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution’, *Genetics*, 169(2), pp. 1157–1164. doi: 10.1534/genetics.104.037051.

Henriques, S. F. *et al.* (2019) ‘Multiple independent L-gulonolactone oxidase (GULO) gene losses and vitamin C synthesis reacquisition events in non-Deuterostomian animal species’, *BMC Evolutionary Biology*, 19(1). doi: 10.1186/s12862-019-1454-8.

Ishigami, T. *et al.* (2001) ‘Regulatory effects of senescence marker protein 30 on the proliferation of hepatocytes’, *Pathology International*, 51(7), pp. 491–497. doi: 10.1046/j.1440-1827.2001.01238.x.

Joppich, C. *et al.* (2009) ‘Umbrea, a chromo shadow domain protein in Drosophila melanogaster heterochromatin, interacts with Hip, HP1 and HOAP’, *Chromosome Research*, 17(1), pp. 19–36. doi: 10.1007/s10577-008-9002-1.

Kaessmann, H., Vinckenbosch, N. and Long, M. (2009) ‘RNA-based gene duplication: mechanistic and evolutionary insights’, *Nature Reviews Genetics*, 10(1), pp. 19–31. doi: 10.1038/nrg2487.

de Koning, H. and Diallinas, G. (2000) ‘Nucleobase transporters’, p. 20.

Kourkoulou, A., Pittis, A. A. and Diallinas, G. (2018) ‘Evolution of substrate specificity in the Nucleobase-Ascorbate Transporter (NAT) protein family’, *Microbial Cell*, 5(6), pp. 280–292. doi: 10.15698/mic2018.06.636.

Krebs, J. E., Goldstein, E. S. and Kilpatrick, S. T. (2018) *Lewin’s genes XII*. Burlington, MA: Jones & Bartlett Learning.

Kreitman, M. and Comeron, J. M. (1999) ‘Coding sequence evolution’, *Current Opinion in Genetics & Development*, 9(6), pp. 637–641. doi: 10.1016/S0959-437X(99)00034-9.

Krissinel, E. and Henrick, K. (2007) ‘Inference of Macromolecular Assemblies from Crystalline State’, *Journal of Molecular Biology*, 372(3), pp. 774–797. doi: 10.1016/j.jmb.2007.05.022.

Krylov, D. M. (2003) ‘Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution’, *Genome Research*, 13(10), pp. 2229–2235. doi: 10.1101/gr.1589103.

Kumar, S., Stecher, G. and Tamura, K. (2016) ‘MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets’, *Molecular Biology and Evolution*, 33(7), pp. 1870–1874. doi: 10.1093/molbev/msw054.

Kuo, S.-M. *et al.* (2004) ‘Gender and Sodium-Ascorbate Transporter Isoforms Determine Ascorbate Concentrations in Mice’, *The Journal of Nutrition*, 134(9), pp. 2216–2221. doi: 10.1093/jn/134.9.2216.

Laurentino, S. S. *et al.* (2012) ‘Regucalcin, a calcium-binding protein with a role in male reproduction?’, *Molecular Human Reproduction*, 18(4), pp. 161–170. doi: 10.1093/molehr/gar075.

Lee, J. H. *et al.* (2006) ‘Immunohistochemical localization of sodium-dependent l-ascorbic acid transporter 1 protein in rat kidney’, *Histochemistry and Cell Biology*, 126(4), pp. 491–494. doi: 10.1007/s00418-006-0186-1.

Lee, S. F. *et al.* (2011) 'Molecular Basis of Adaptive Shift in Body Size in Drosophila melanogaster: Functional and Sequence Analyses of the Dca Gene', *Molecular Biology and Evolution*, 28(8), pp. 2393–2402. doi: 10.1093/molbev/msr064.

Li, H. *et al.* (2017) 'Mitochondrial phylogenomics of Hemiptera reveals adaptive innovations driving the diversification of true bugs', *Proceedings of the Royal Society B: Biological Sciences*, 284(1862), p. 20171223. doi: 10.1098/rspb.2017.1223.

Liang, H. and Li, W.-H. (2007) 'Gene essentiality, gene duplicability and protein connectivity in human and mouse', *Trends in Genetics*, 23(8), pp. 375–378. doi: 10.1016/j.tig.2007.04.005.

Liao, B.-Y. and Zhang, J. (2007) 'Mouse duplicate genes are as essential as singletons', *Trends in Genetics*, 23(8), pp. 378–381. doi: 10.1016/j.tig.2007.05.006.

Loewus, F. (1999) 'Biosynthesis and metabolism of ascorbic acid in plants and of analogs of ascorbic acid in fungi', *Phytochemistry*, 52(2), pp. 193–210. doi: 10.1016/S0031-9422(99)00145-4.

Long, M. *et al.* (2003) 'The origin of new genes: glimpses from the young and old', *Nature Reviews Genetics*, 4(11), pp. 865–875. doi: 10.1038/nrg1204.

Lynch, M. *et al.* (2001) 'The Probability of Preservation of a Newly Arisen Gene Duplicate', p. 16.

Maddison, D. R., Schulz, K.-S. and Maddison, W. P. (2007) 'The Tree of Life Web Project', p. 22.

Maia, C. *et al.* (2009) 'Regucalcin is under-expressed in human breast and prostate cancers: Effect of sex steroid hormones', *Journal of Cellular Biochemistry*, 107(4), pp. 667–676. doi: 10.1002/jcb.22158.

Makino, T., Hokamp, K. and McLysaght, A. (2009) 'The complex relationship of gene duplication and essentiality', p. 4.

Marques, R. *et al.* (2014) 'The diverse roles of calcium-binding protein regucalcin in cell biology: from tissue expression and signalling to disease', *Cellular and Molecular Life Sciences*, 71(1), pp. 93–111. doi: 10.1007/s00018-013-1323-3.

McKechnie, S. W. *et al.* (2010) 'A clinally varying promoter polymorphism associated with adaptive variation in wing size in *Drosophila*: *DROSOPHILA* : SIZE VARIATION AND THE *DCA* PROMOTER', *Molecular Ecology*, 19(4), pp. 775–784. doi: 10.1111/j.1365-294X.2009.04509.x.

Miklos, G. L. G. and Rubin, G. M. (1996) 'The Role of the Genome Project in Determining Gene Function: Insights from Model Organisms', *Cell*, 86(4), pp. 521–529. doi: 10.1016/S0092-8674(00)80126-9.

Misawa, H. and Yamaguchi, M. (2000) 'The gene of Ca2+-binding protein regucalcin is highly conserved in vertebrate species.', *International Journal of Molecular Medicine*. doi: 10.3892/ijmm.6.2.191.

Misof, B. *et al.* (2014) 'Phylogenomics resolves the timing and pattern of insect evolution', *Science*, 346(6210), pp. 763–767. doi: 10.1126/science.1257570.

Muñoz, A. *et al.* (2015) 'Cis-regulatory elements involved in species-specific transcriptional regulation of the SVCT1 gene in rat and human hepatoma cells', *Free Radical Biology and Medicine*, 85, pp. 183–196. doi: 10.1016/j.freeradbiomed.2015.04.024.

Murrell, B. *et al.* (2013) 'FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection', *Molecular Biology and Evolution*, 30(5), pp. 1196–1205. doi: 10.1093/molbev/mst030.

Nakajima, M., Murata, T. and Yamaguchi, M. (1999) 'Expression of calcium-binding protein regucalcin mRNA in the cloned rat hepatoma cells (H4-II-E) is stimulated through Ca2+ signaling factors: Involvement of protein kinase C', p. 8.

Nikapitiya, C. *et al.* (2008) 'Molecular characterization and expression analysis of regucalcin in disk abalone (Haliotis discus discus): Intramuscular calcium administration stimulates the regucalcin mRNA expression', *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 150(1), pp. 117–124. doi: 10.1016/j.cbpb.2008.02.004.

Notredame, C., Higgins, D. G. and Heringa, J. (2000) 'T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton', *Journal of Molecular Biology*, 302(1), pp. 205–217. doi: 10.1006/jmbi.2000.4042.

Ohno, S. (2014) *Evolution by Gene Duplication*. Berlin: Springer Berlin.

Patananan, A. N. *et al.* (2015) 'The invertebrate Caenorhabditis elegans biosynthesizes ascorbate', *Archives of Biochemistry and Biophysics*, 569, pp. 32–44. doi: 10.1016/j.abb.2015.02.002.

Reboiro-Jato, D. *et al.* (2012) 'ADOPS - Automatic Detection Of Positively Selected Sites', *Journal of Integrative Bioinformatics*, 9(3). doi: 10.1515/jib-2012-200.

Ronquist, F. *et al.* (2012) 'MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space', *Systematic Biology*, 61(3), pp. 539–542. doi: 10.1093/sysbio/sys029.

Roy, A., Kucukural, A. and Zhang, Y. (2010) 'I-TASSER: a unified platform for automated protein structure and function prediction', *Nature Protocols*, 5(4), pp. 725–738. doi: 10.1038/nprot.2010.5.

Scott, S. H. and Bahnson, B. J. (2011) 'Senescence marker protein 30: functional and structural insights to its unknown physiological function', *BioMolecular Concepts*, 2(6). doi: 10.1515/BMC.2011.041.

Shigeoka, S., Nakano, Y. and Kitaoka, S. (1979) 'The biosynthetic pathway of L-ascorbic acid in Euglena gracilis Z.', *Journal of Nutritional Science and Vitaminology*, 25(4), pp. 299–307. doi: 10.3177/jnsv.25.299.

Shimokawa, N. and Yamaguchi, M. (1993) 'Molecular cloning and sequencing of the cDNA coding for a calcium-binding protein regucalcin from rat liver', *FEBS Letters*, 327(3), pp. 251–255. doi: 10.1016/0014-5793(93)80998-A.

Smirnoff, N. (2018) 'Ascorbic acid metabolism and functions: A comparison of plants and mammals', *Free Radical Biology and Medicine*, 122, pp. 116–129. doi: 10.1016/j.freeradbiomed.2018.03.033.

Song, N., Liang, A.-P. and Bu, C.-P. (2012) 'A Molecular Phylogeny of Hemiptera Inferred from Mitochondrial Genome Sequences', *PLoS ONE*. Edited by S. Ho, 7(11), p. e48778. doi: 10.1371/journal.pone.0048778.

Su, Z. and Gu, X. (2008) 'Predicting the Proportion of Essential Genes in Mouse Duplicates Based on Biased Mouse Knockout Genes', *Journal of Molecular Evolution*, 67(6), pp. 705–709. doi: 10.1007/s00239-008-9170-9.

Tamura, K. *et al.* (2007) 'MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0', *Molecular Biology and Evolution*, 24(8), pp. 1596–1599. doi: 10.1093/molbev/msm092.

Vázquez, N. *et al.* (2019) 'EvoPPI 1.0: a Web Platform for Within- and Between-Species Multiple Interactome Comparisons and Application to Nine PolyQ Proteins Determining Neurodegenerative Diseases', *Interdisciplinary Sciences: Computational Life Sciences*, 11(1), pp. 45–56. doi: 10.1007/s12539-019-00317-y.

Vierstraete, E. *et al.* (2003) 'Proteomics in Drosophila melanogaster: first 2D database of larval hemolymph proteins', *Biochemical and Biophysical Research Communications*, 304(4), pp. 831–838. doi: 10.1016/S0006-291X(03)00683-1.

de Vries, S. J. and Bonvin, A. M. J. J. (2011) 'CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK', *PLoS ONE*. Edited by N. Fernandez-Fuentes, 6(3), p. e17695. doi: 10.1371/journal.pone.0017695.

Walsh, J. B. (1995) 'How Often Do Duplicated Genes Evolve New Functions?', *Gene Duplication*, p. 8.

Wheeler, G. *et al.* (2015) 'Evolution of alternative biosynthetic pathways for vitamin C following plastid acquisition in photosynthetic eukaryotes', *eLife*, 4. doi: 10.7554/eLife.06369.

Wheeler, G. L., Jones, M. A. and Smirnoff, N. (1998) 'The biosynthetic pathway of vitamin C in higher plants', 393, p. 5.

Wilson, A. C., Carlson, S. S. and White, T. J. (1977) 'Biochemical Evolution', p. 67.

Yamaguchi, M. (2011) 'The transcriptional regulation of regucalcin gene expression', *Molecular and Cellular Biochemistry*, 346(1–2), pp. 147–171. doi: 10.1007/s11010-010-0601-8.

Yamaguchi, M. (2013) 'Role of regucalcin in cell nuclear regulation: involvement as a transcription factor', *Cell and Tissue Research*, 354(2), pp. 331–341. doi: 10.1007/s00441-013-1665-z.

Yamaguchi, M. *et al.* (2016) 'Prolonged survival in hepatocarcinoma patients with increased regucalcin gene expression: HepG2 cell proliferation is suppressed by overexpression of regucalcin in vitro', *International Journal of Oncology*, 49(4), pp. 1686–1694. doi: 10.3892/ijo.2016.3669.

Yamaguchi, M. and Murata, T. (2013) 'Involvement of regucalcin in lipid metabolism and diabetes', *Metabolism*, 62(8), pp. 1045–1051. doi: 10.1016/j.metabol.2013.01.023.

Yamaguchi, M. and Yamamoto, T. (1978) 'Purification of calcium binding substance from soluble fraction of normal rat liver.', *CHEMICAL & PHARMACEUTICAL BULLETIN*, 26(6), pp. 1915–1918. doi: 10.1248/cpb.26.1915.

Yamamoto, S. *et al.* (2010) 'Identification and Functional Characterization of the First Nucleobase Transporter in Mammals: IMPLICATION IN THE SPECIES DIFFERENCE IN THE INTESTINAL ABSORPTION MECHANISM OF NUCLEOBASES AND THEIR ANALOGS BETWEEN

HIGHER PRIMATES AND OTHER MAMMALS', *Journal of Biological Chemistry*, 285(9), pp. 6522–6531. doi: 10.1074/jbc.M109.032961.

Yang, J. *et al.* (2015) 'The I-TASSER Suite: protein structure and function prediction', *Nature Methods*, 12(1), pp. 7–8. doi: 10.1038/nmeth.3213.

Yang, Z. (2007) 'PAML 4: Phylogenetic Analysis by Maximum Likelihood', *Molecular Biology and Evolution*, 24(8), pp. 1586–1591. doi: 10.1093/molbev/msm088.

Zhang, J. (2003) 'Evolution by gene duplication: an update', *Trends in Ecology & Evolution*, 18(6), pp. 292–298. doi: 10.1016/S0169-5347(03)00033-8.

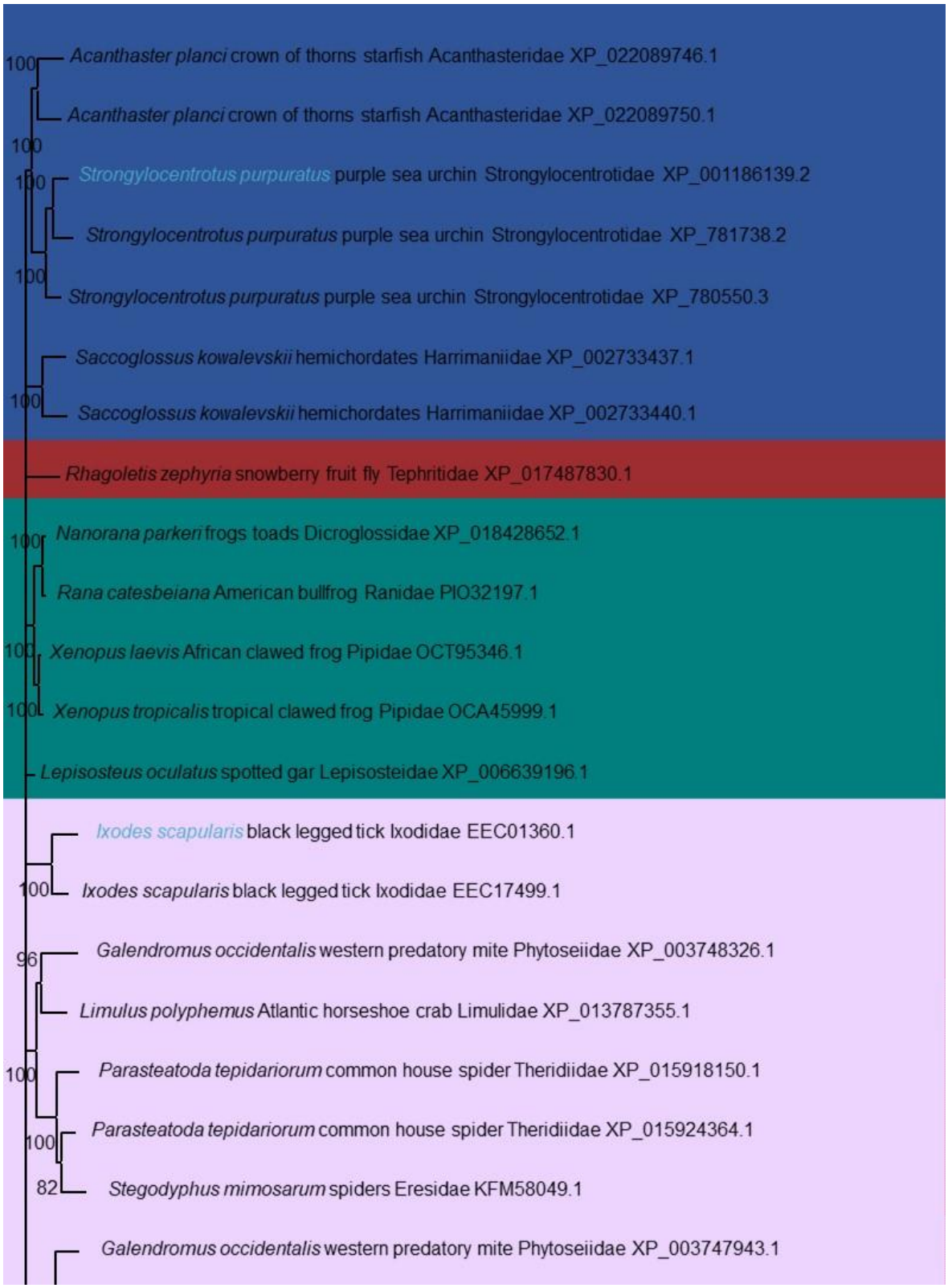
Zhang, S.-Q. *et al.* (2018) 'Evolutionary history of Coleoptera revealed by extensive sampling of genes and species', *Nature Communications*, 9(1). doi: 10.1038/s41467-017-02644-4.

Zhang, Y. (2008) 'I-TASSER server for protein 3D structure prediction', *BMC Bioinformatics*, 9(1). doi: 10.1186/1471-2105-9-40.

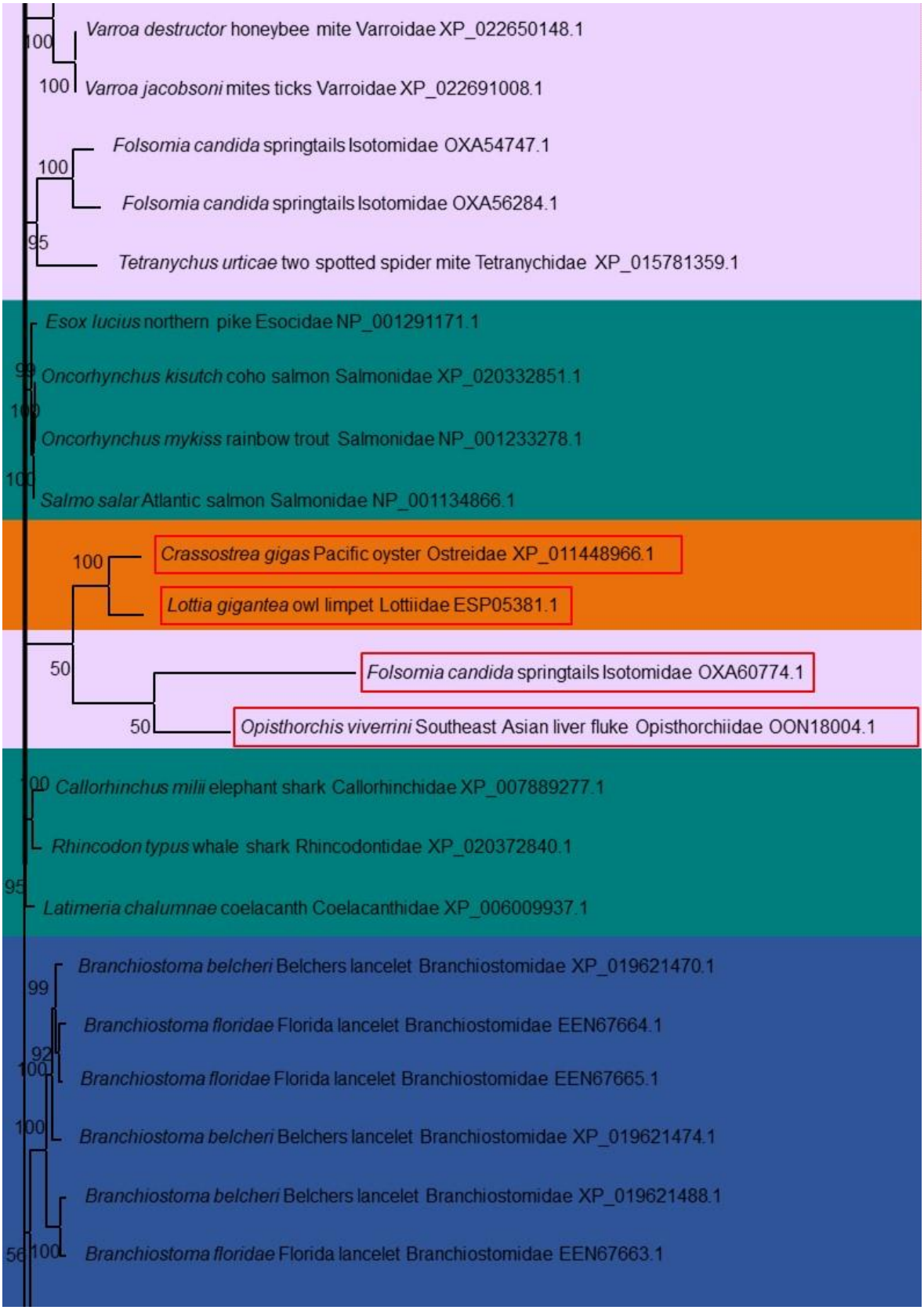
van Zundert, G. C. P. *et al.* (2016) 'The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes', *Journal of Molecular Biology*, 428(4), pp. 720–725. doi: 10.1016/j.jmb.2015.09.014.

Supplementary material:

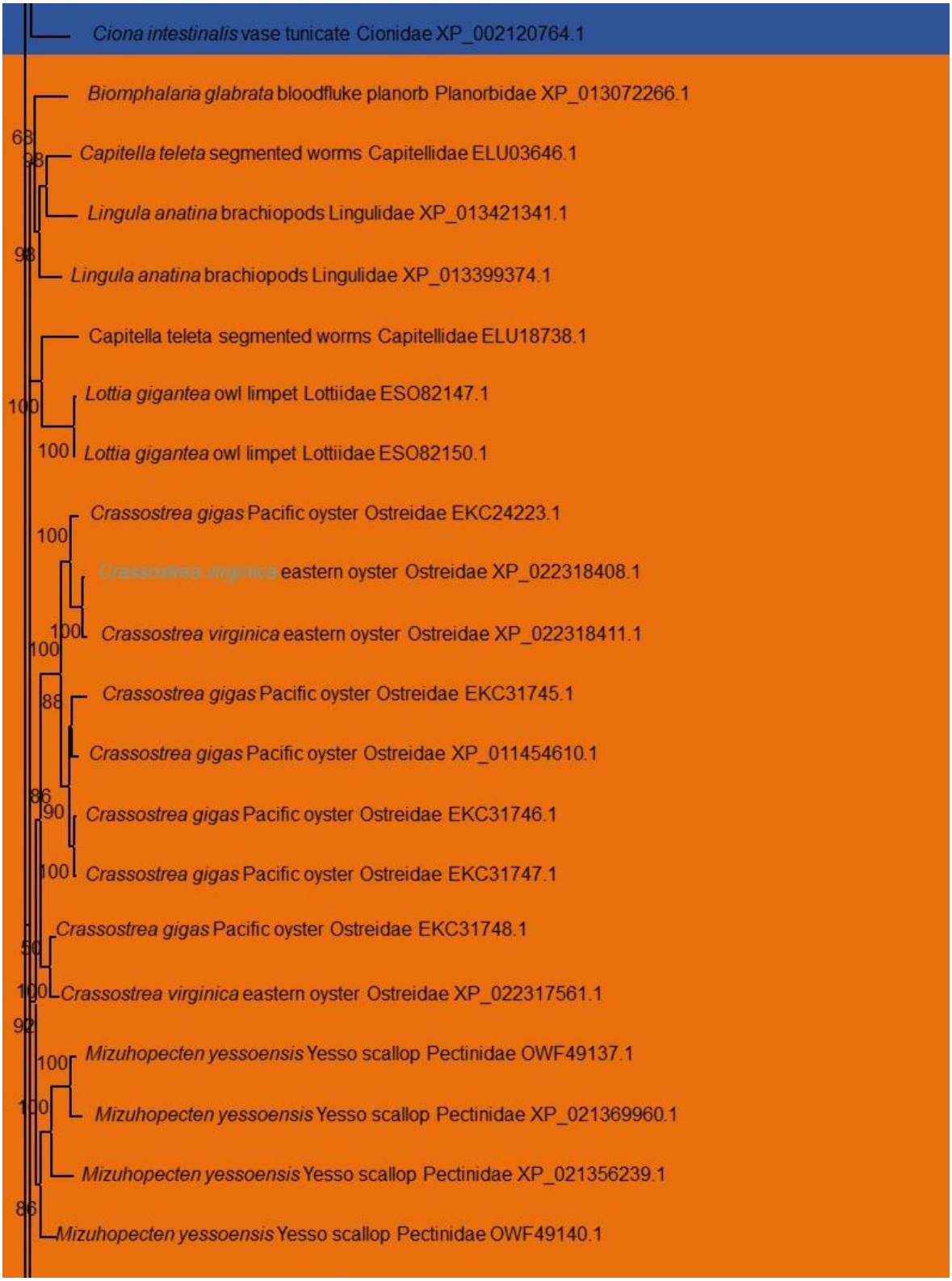
1. Animal Regucalcin phylogenies



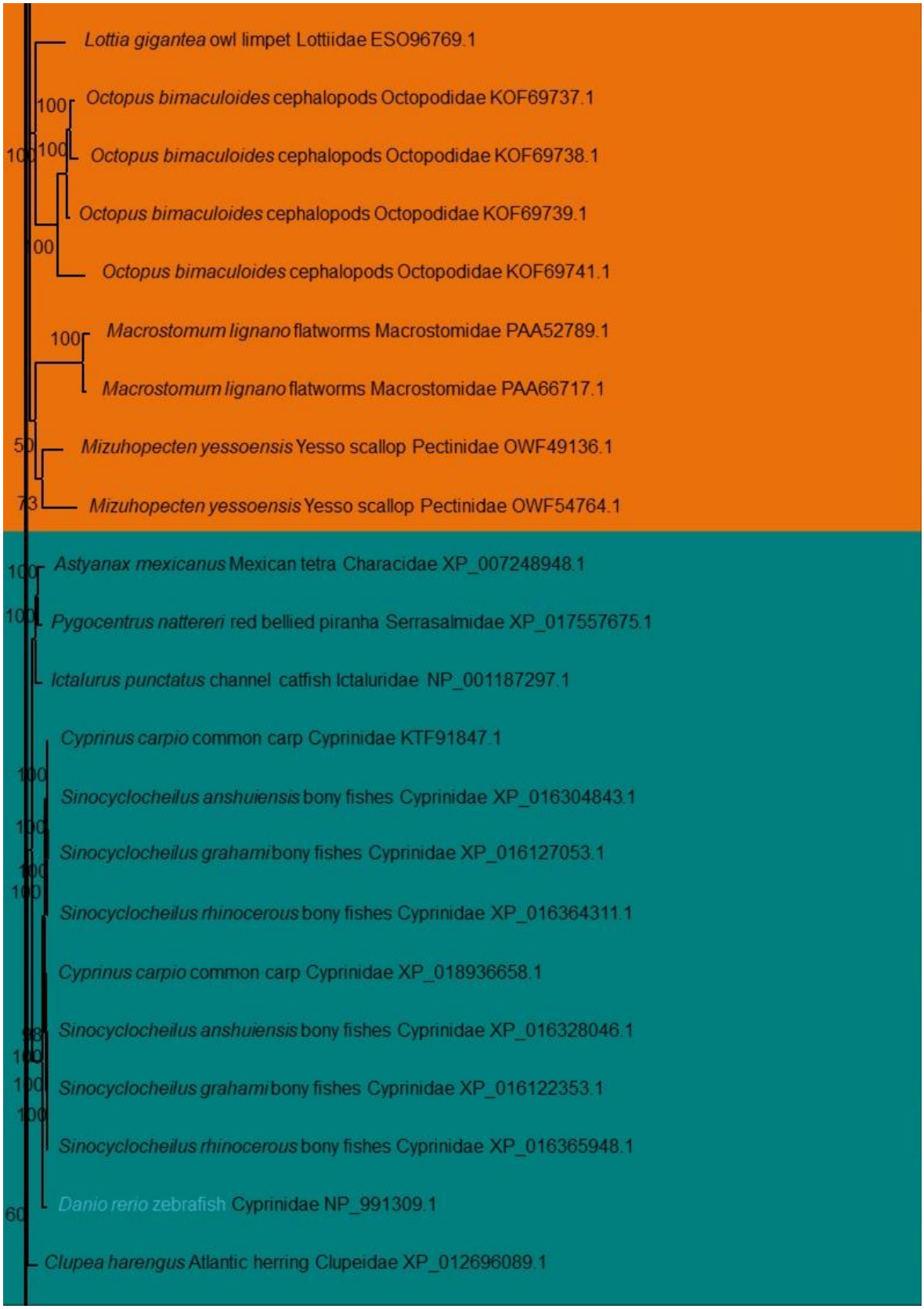
(The image continues on the next page)



(The image continues on the next page)

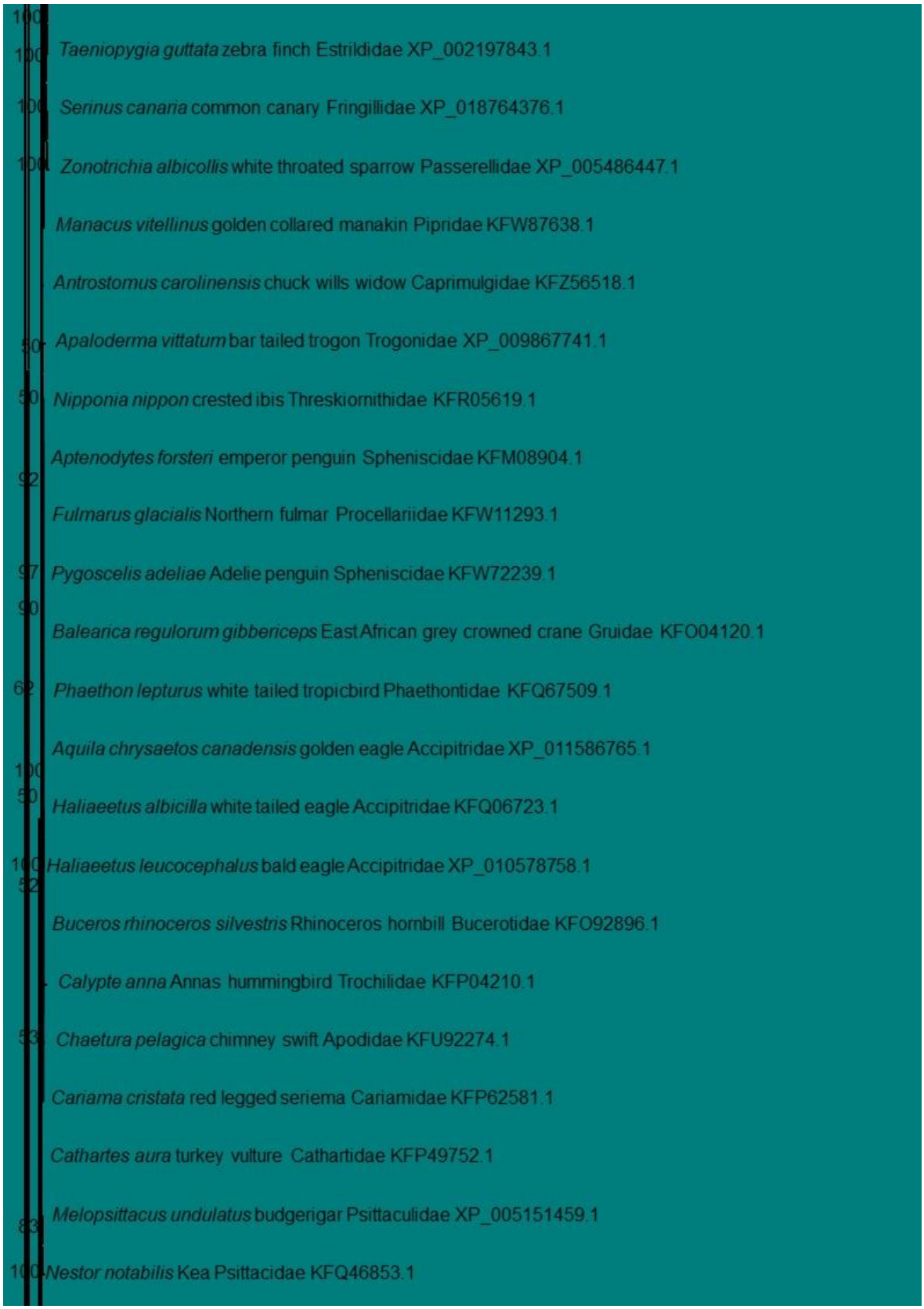


(The image continues on the next page)



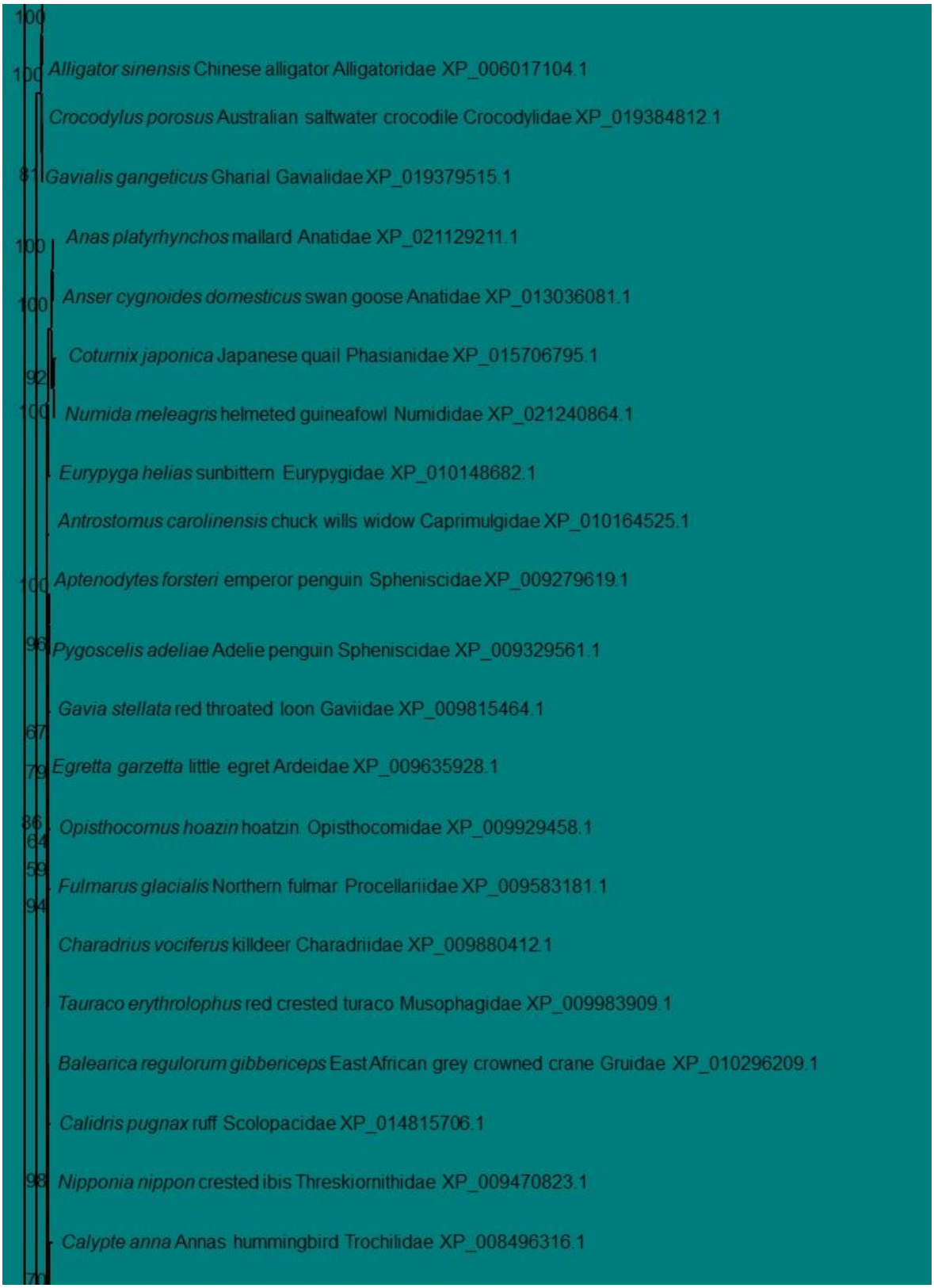
(The image continues on the next page)

(The image continues on the next page)

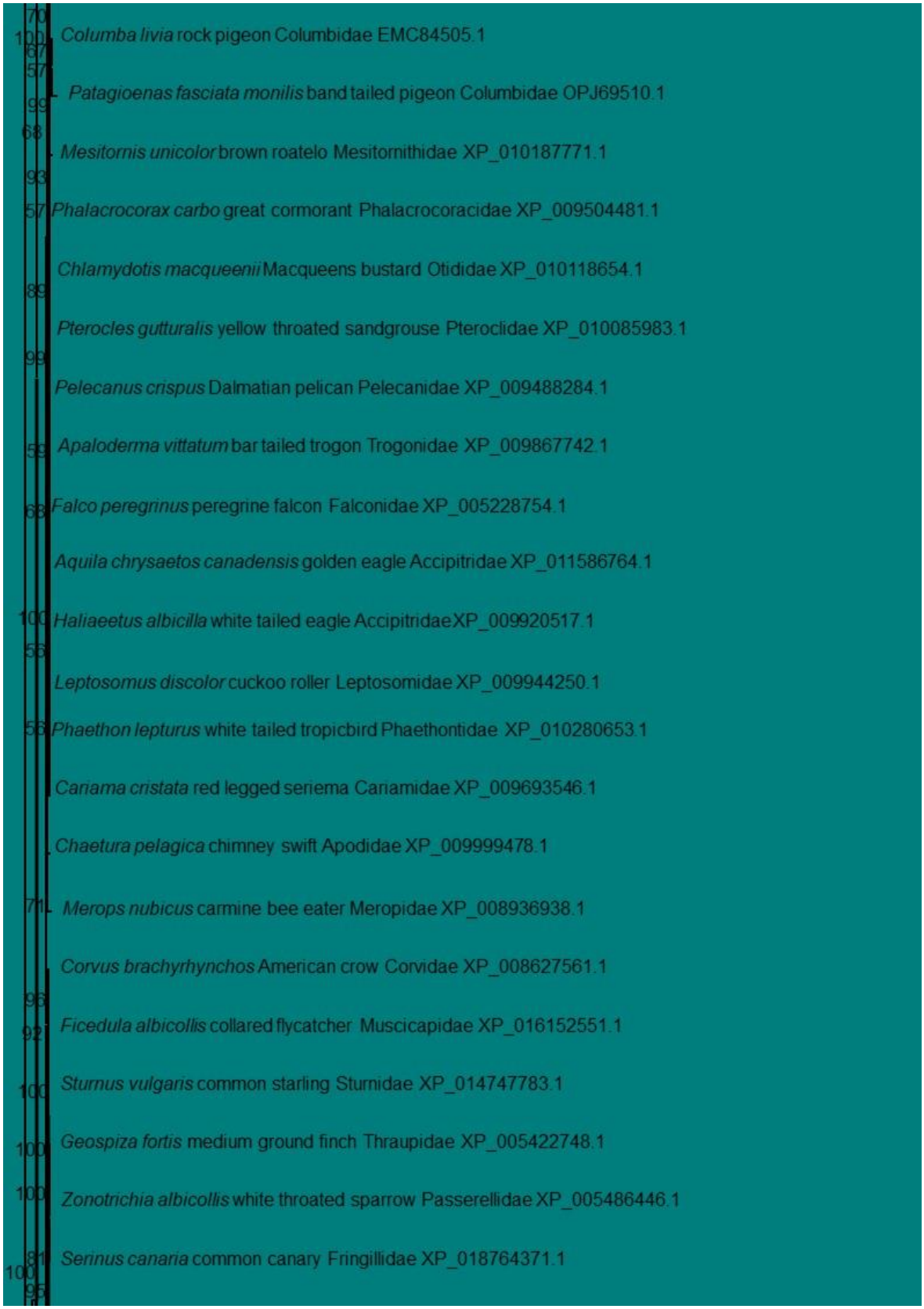


(The image continues on the next page)

(The image continues on the next page)

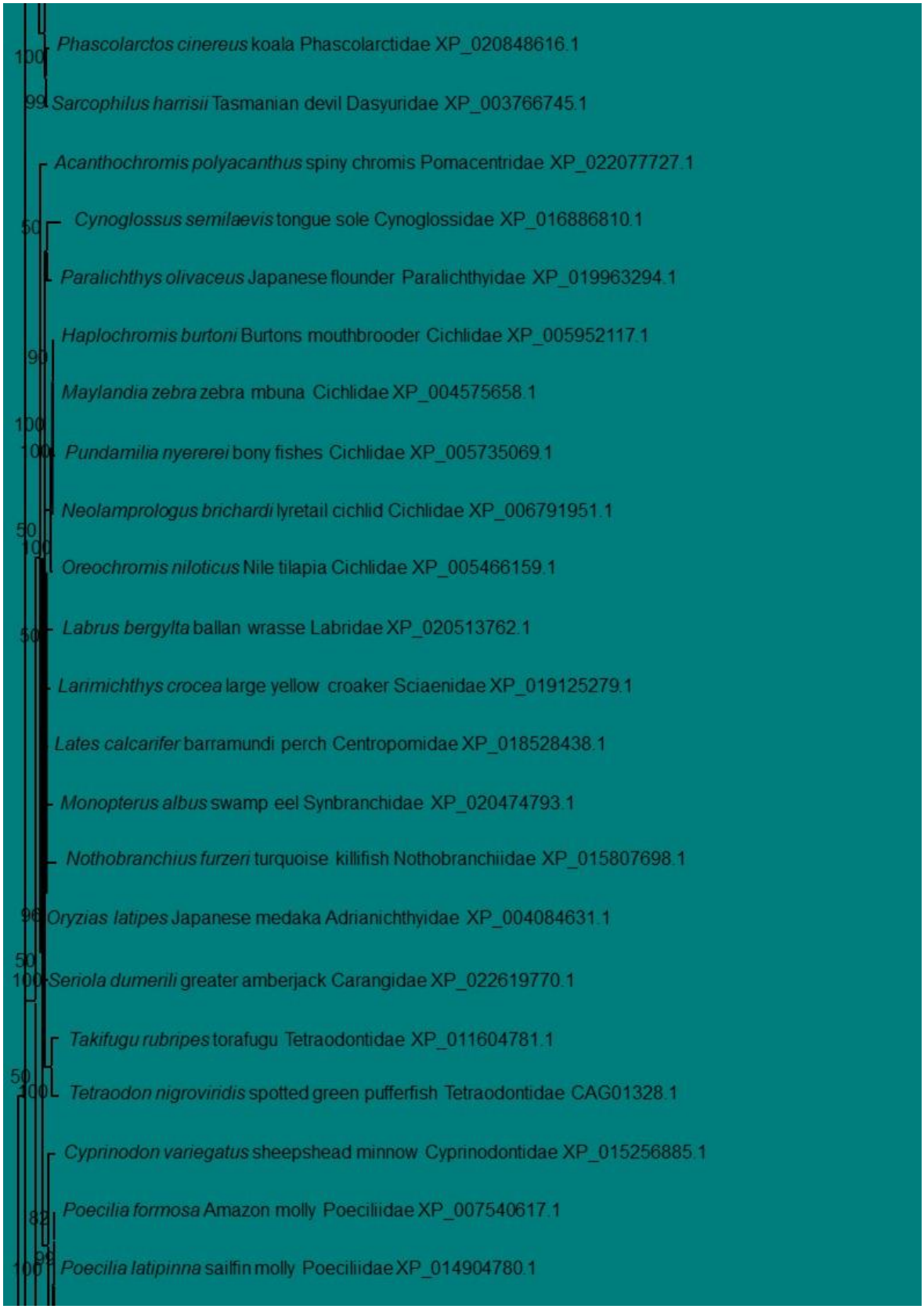


(The image continues on the next page)

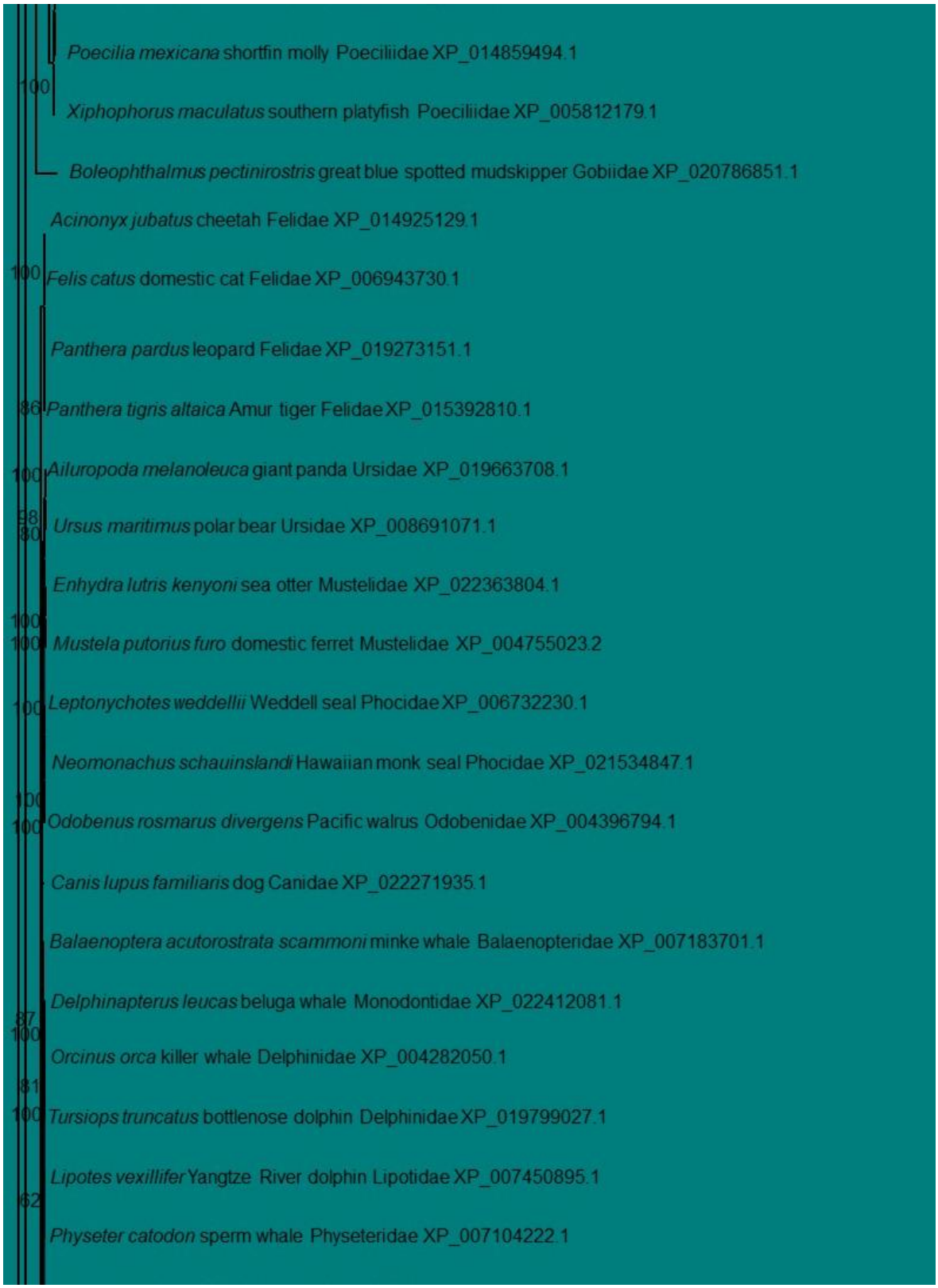


(The image continues on the next page)

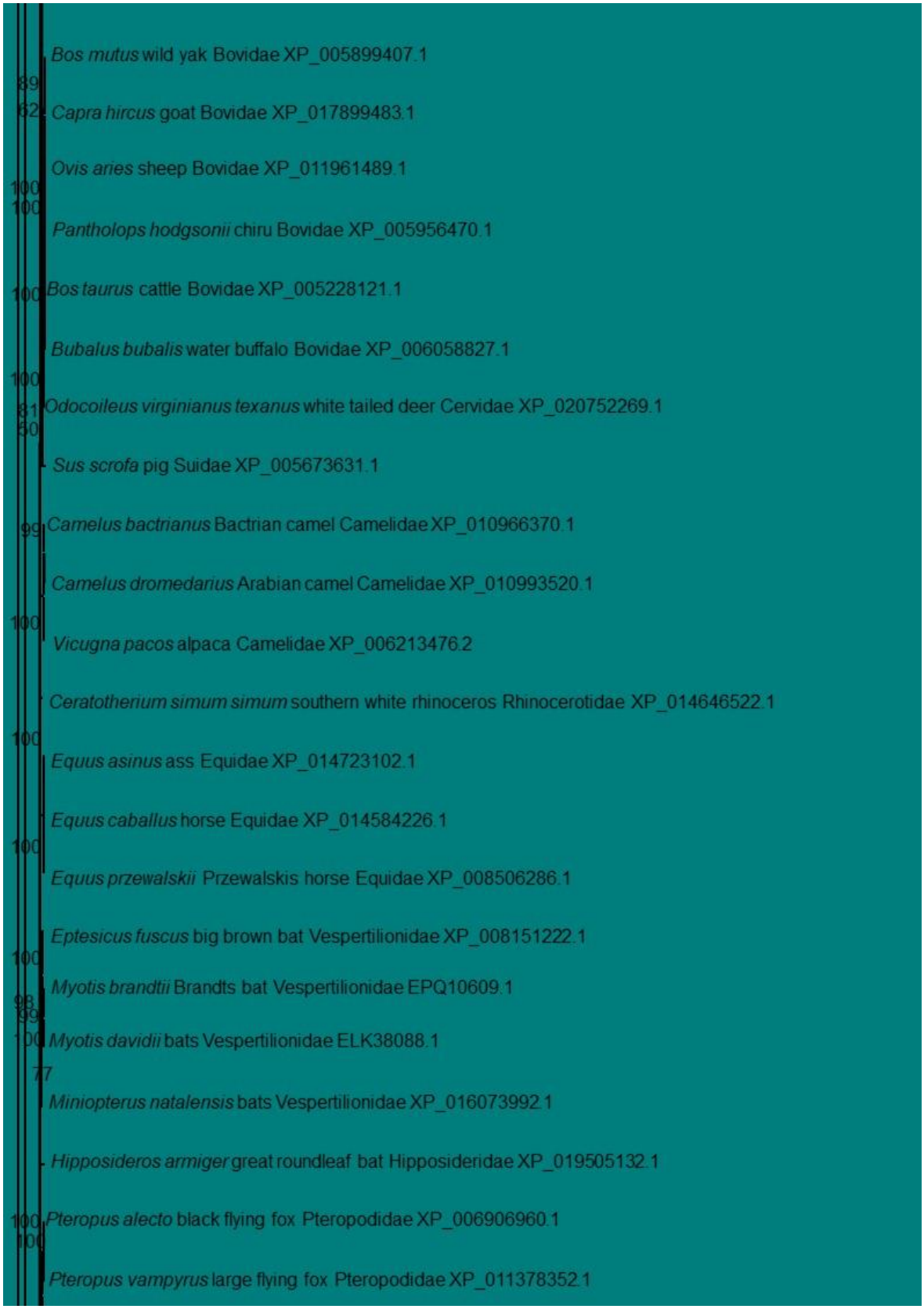
(The image continues on the next page)



(The image continues on the next page)

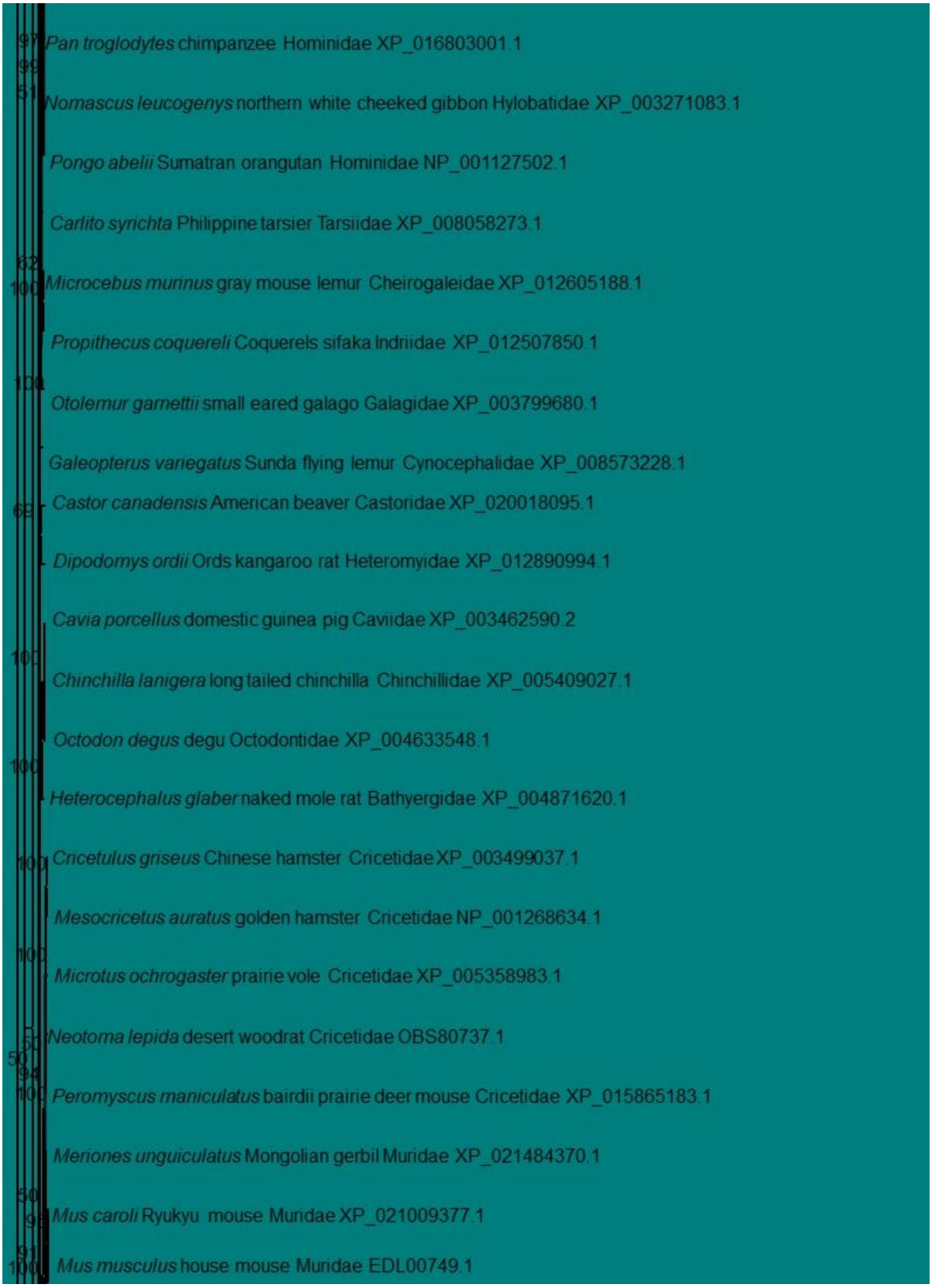


(The image continues on the next page)

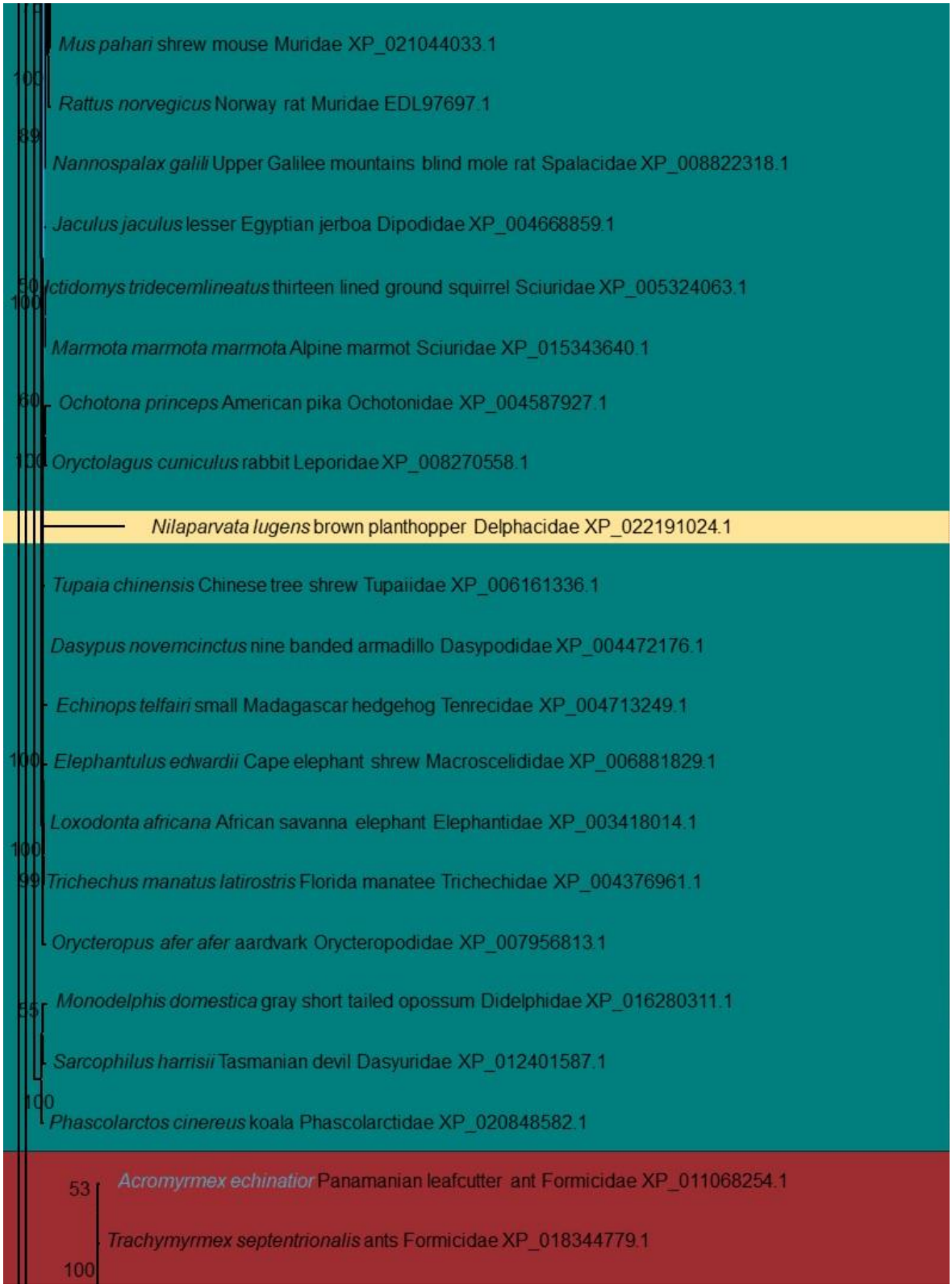


(The image continues on the next page)

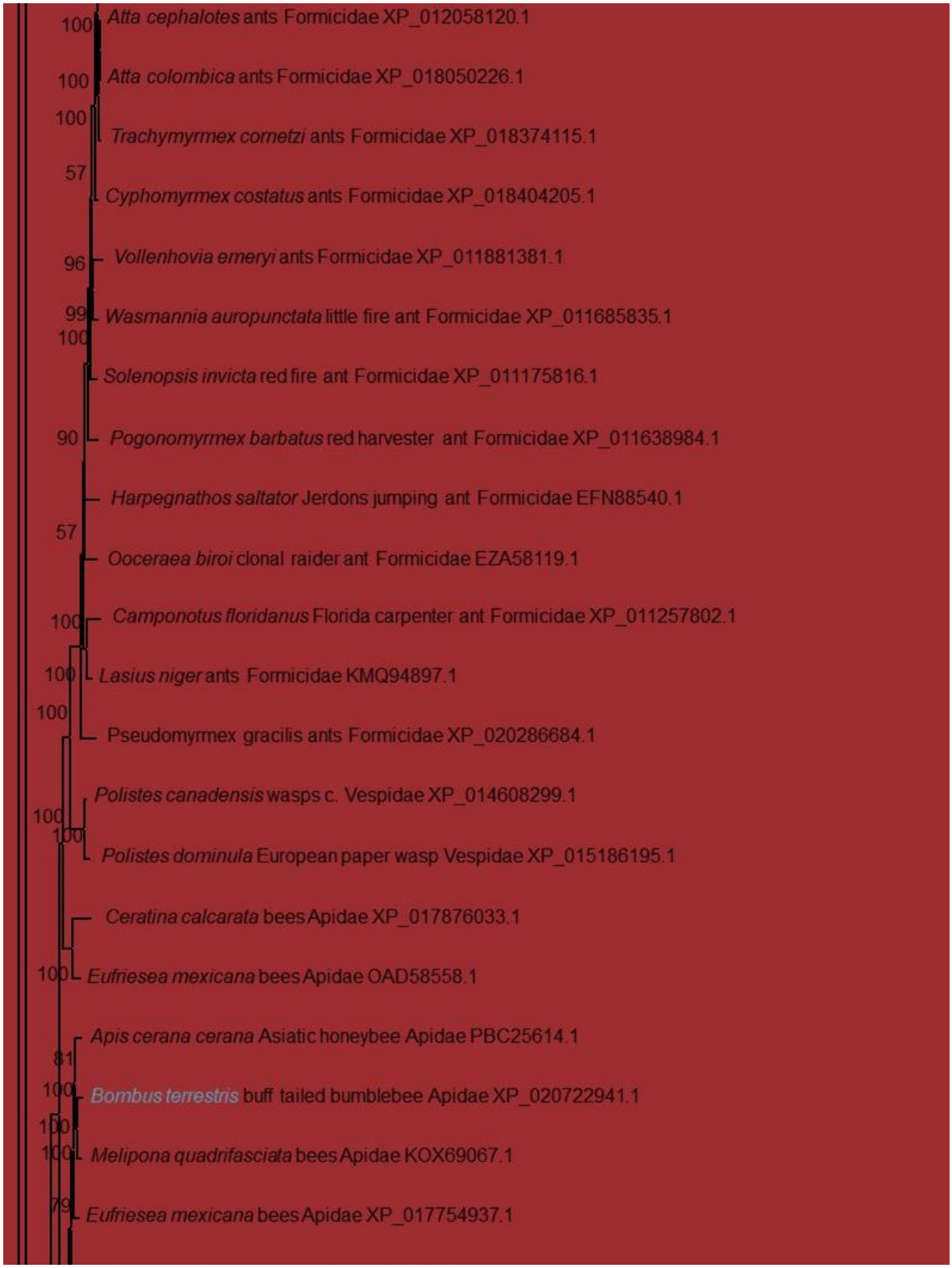
(The image continues on the next page)



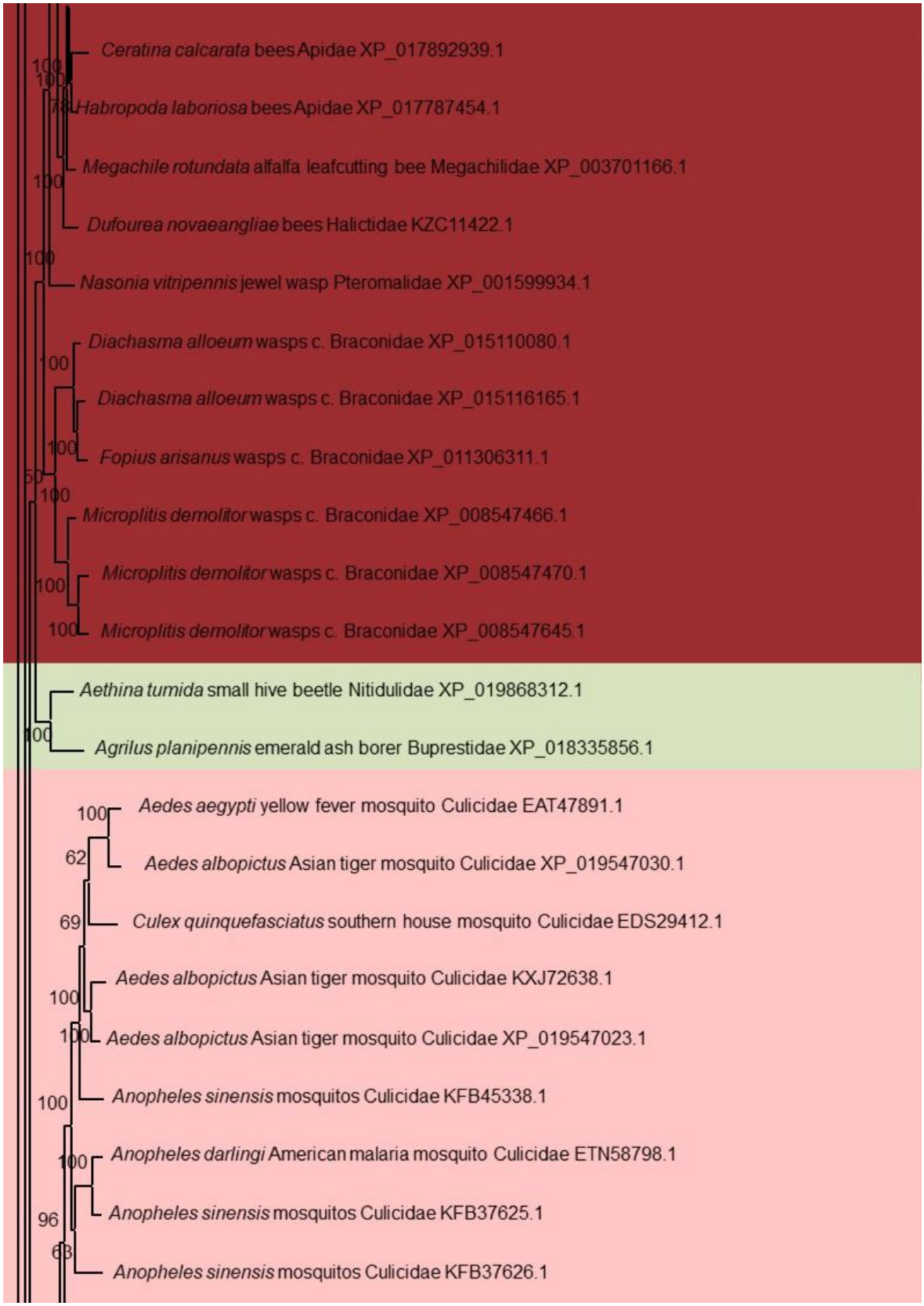
(The image continues on the next page)



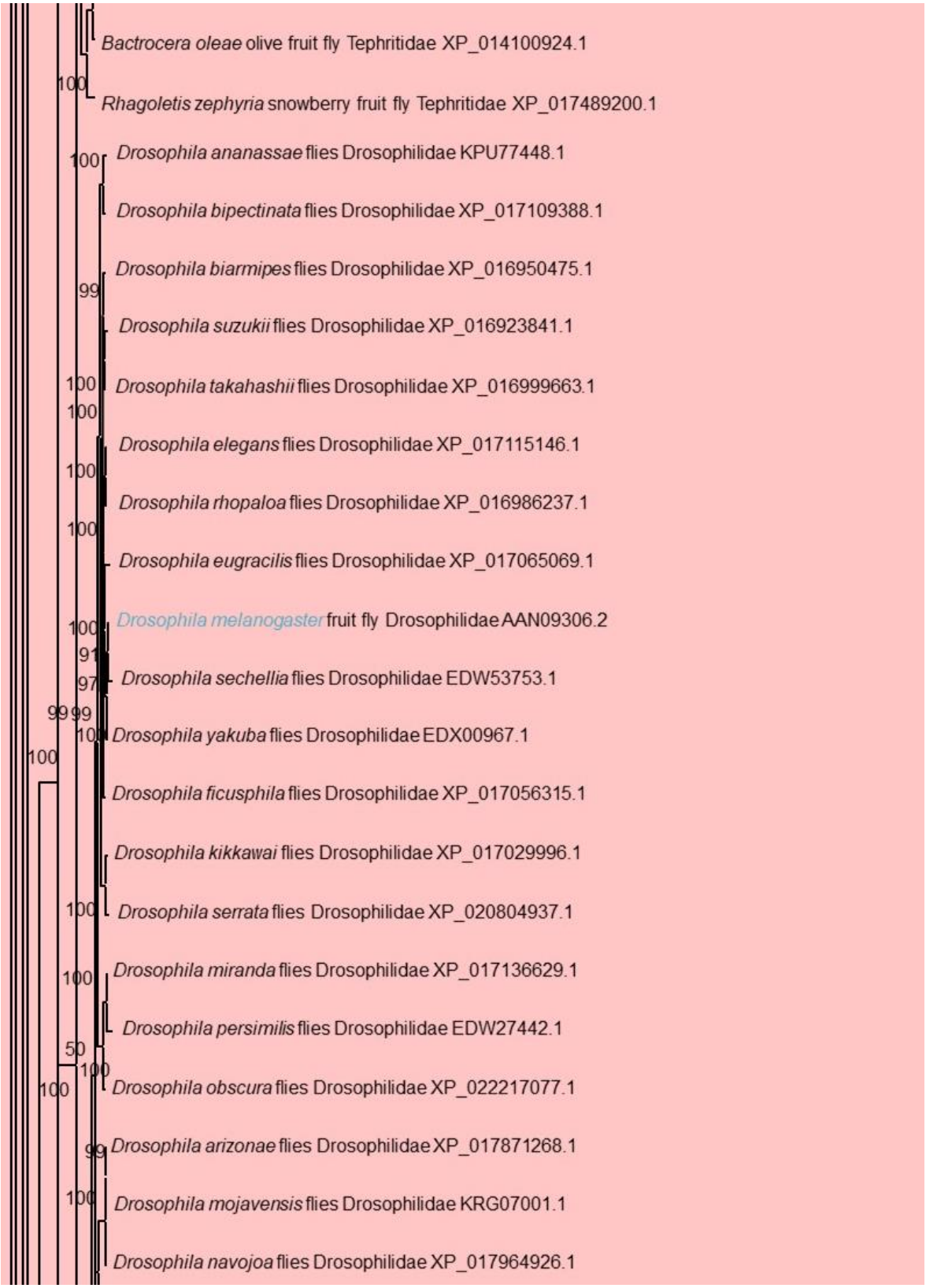
(The image continues on the next page)



(The image continues on the next page)

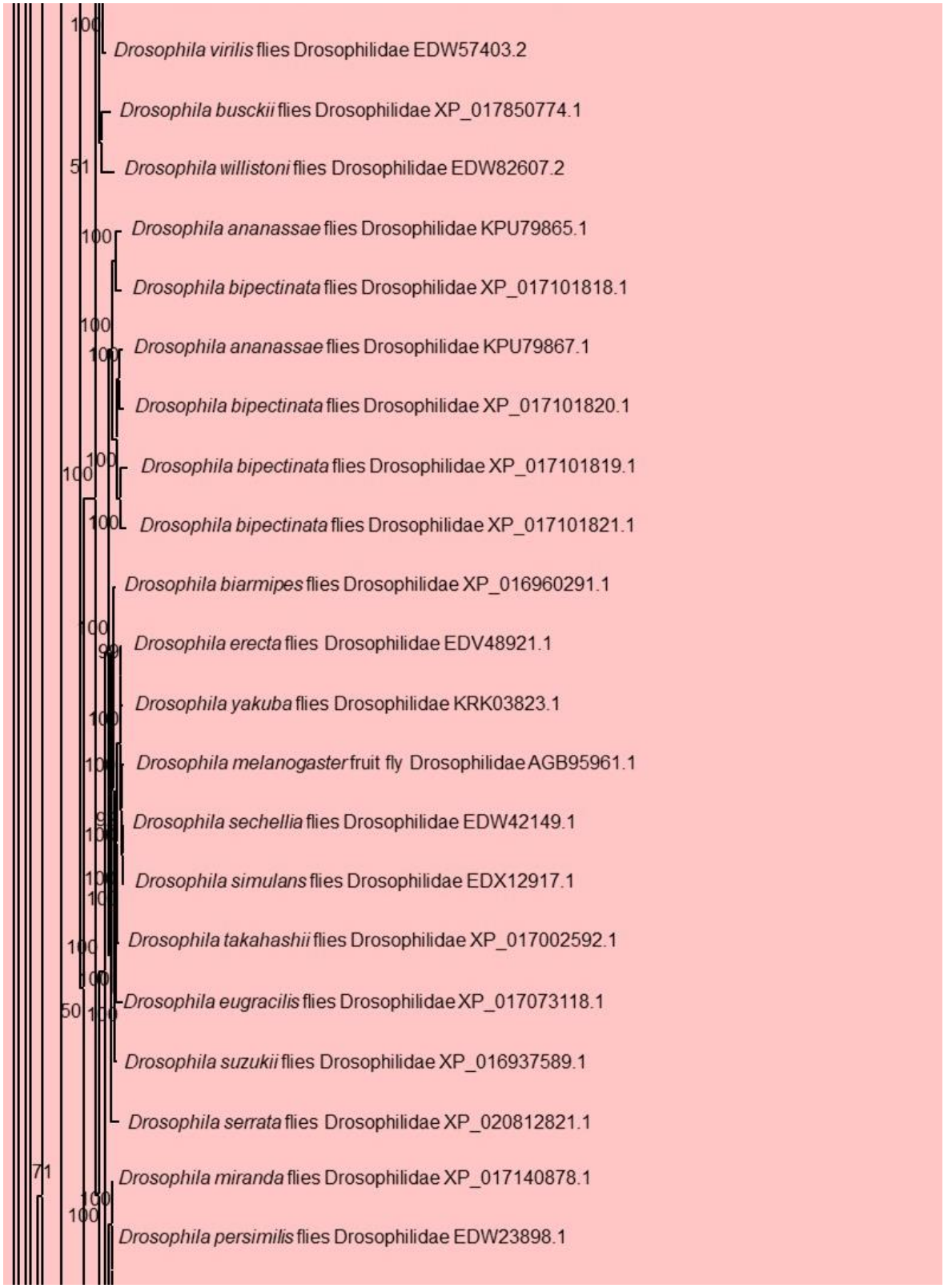


(The image continues on the next page)

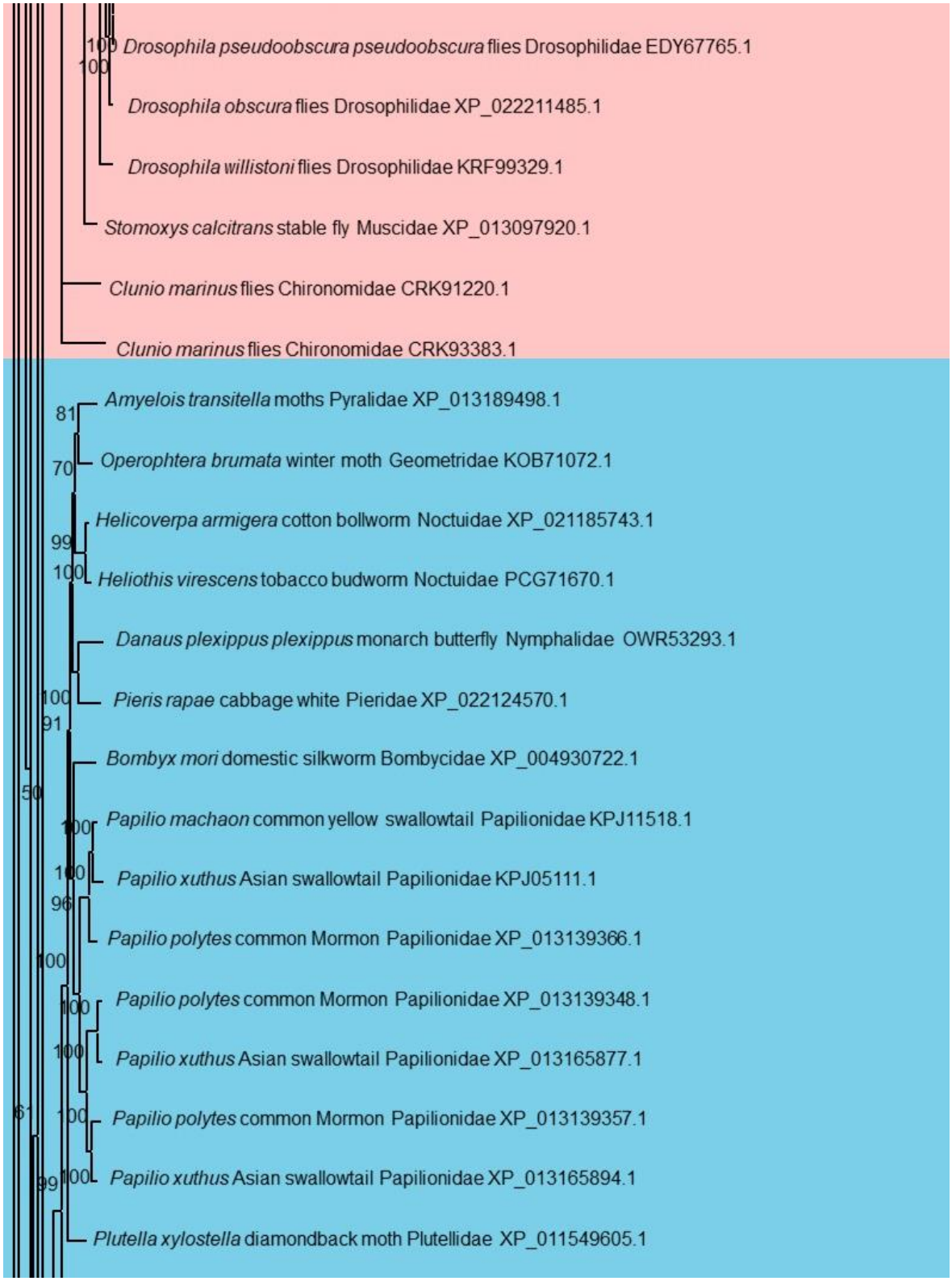


(The image continues on the next page)

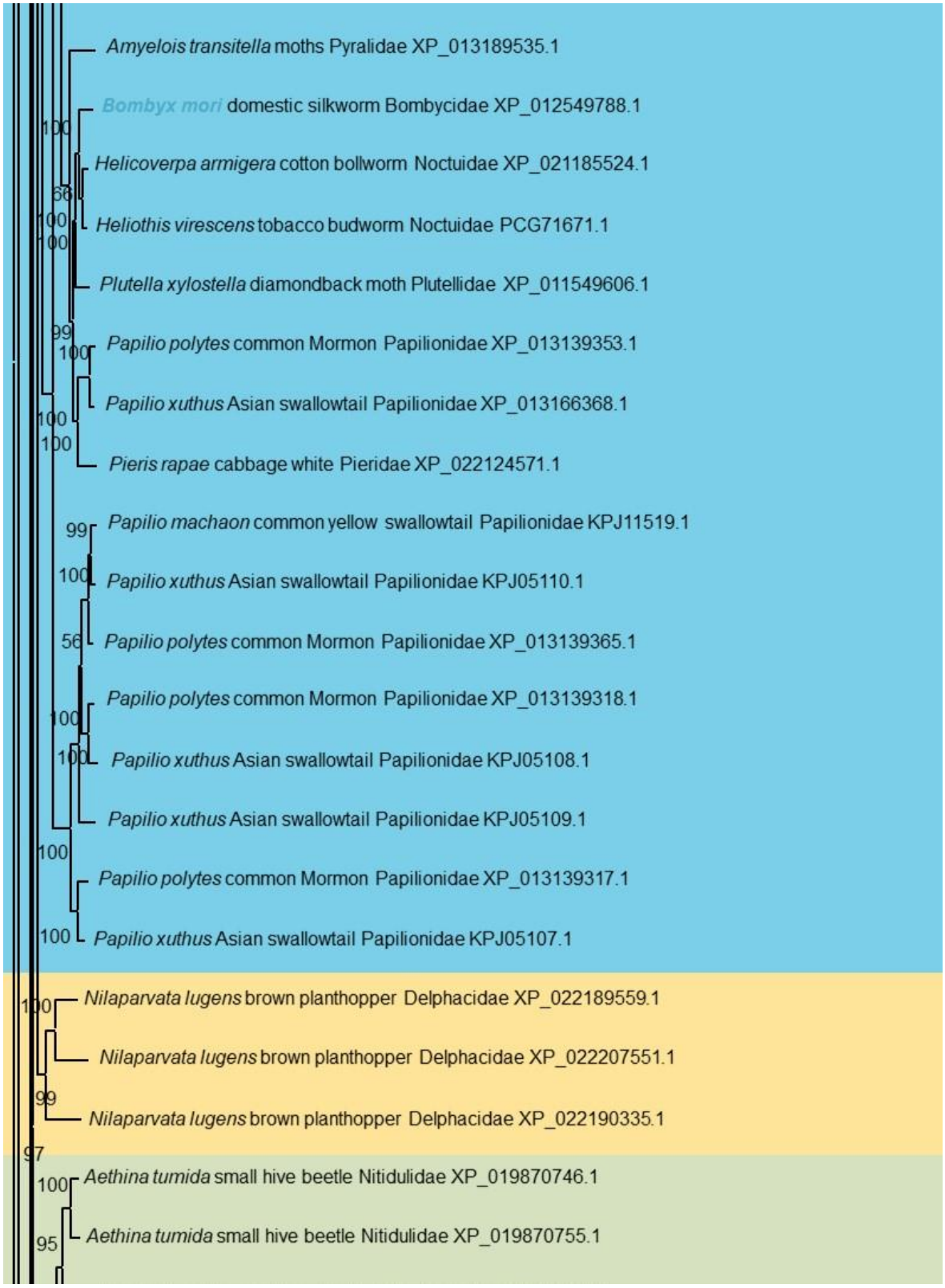
(The image continues on the next page)



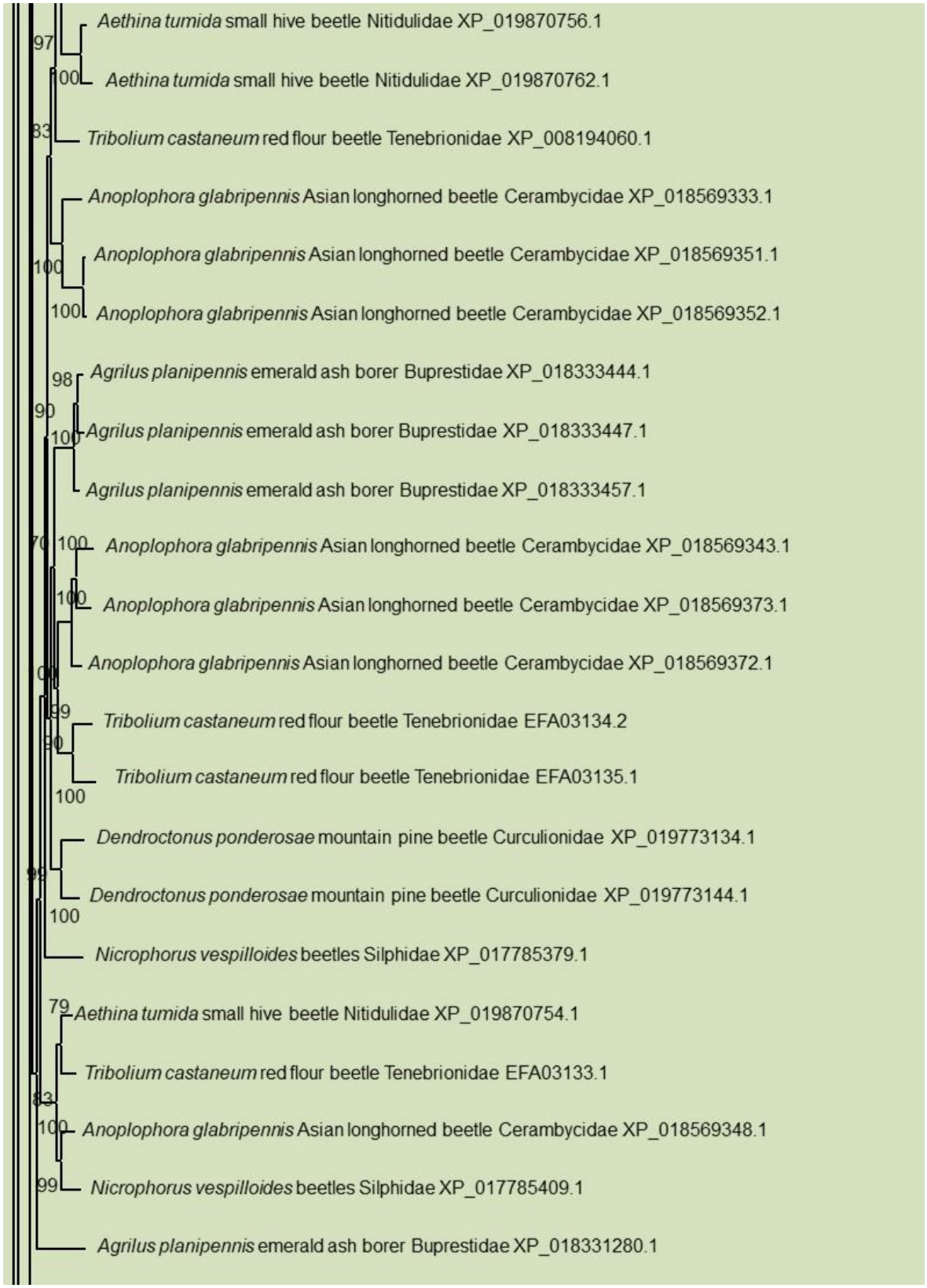
(The image continues on the next page)



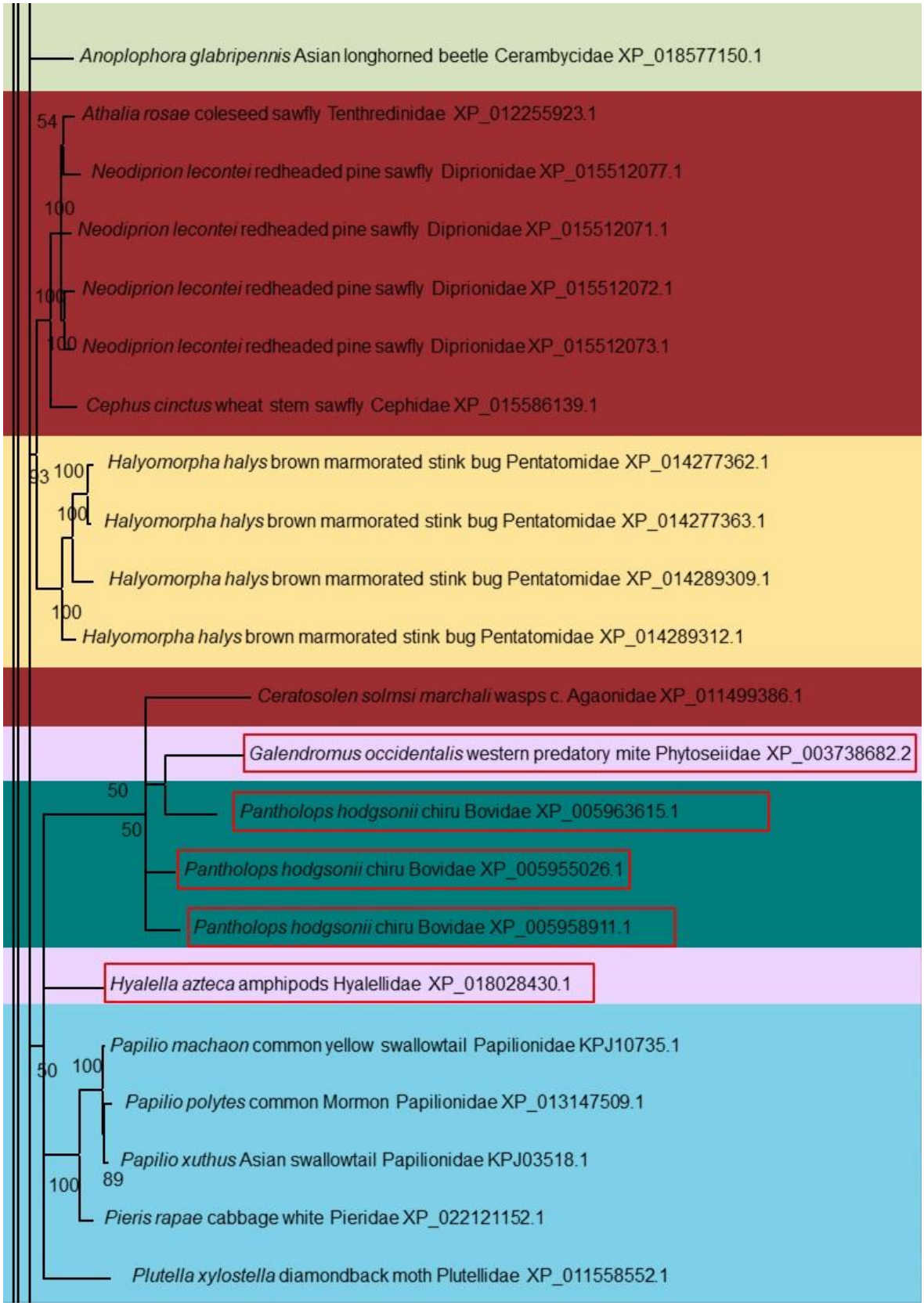
(The image continues on the next page)



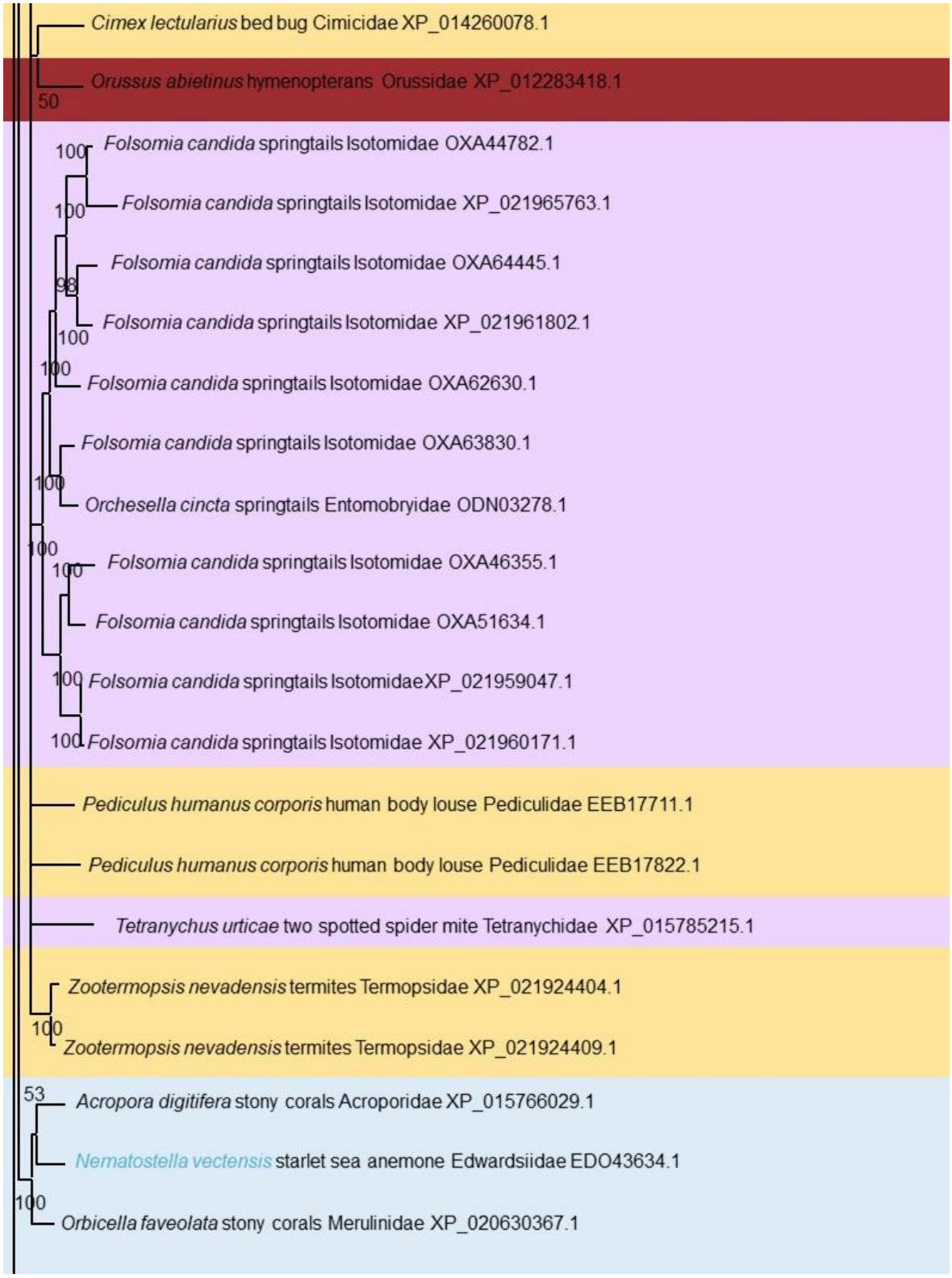
(The image continues on the next page)



(The image continues on the next page)



(The image continues on the next page)



(The image continues on the next page)

Figure S 1 – Bayesian phylogeny of the evolution of the *Regucalcin* gene in the animal kingdom.

Table 13 - The sequences represented in the final phylogeny showing phylogenetic information on more than one species.

Replaced species	Protein Accession number	Replacing species	Protein Accession number
Haliaeetus_leucocephalus	XP_010578757.1		
		Haliaeetus_albicilla	XP_009920517.1
Haliaeetus_albicilla	KFQ06724.1		
Corvus_brachyrhynchos	XP_017584063.1		
Corvus_cornix_cornix	XP_019138371.1		
Corvus_cornix_cornix	XP_019138372.1	*Corvus_brachyrhynchos*	XP_008627561.1
Corvus_cornix_cornix	XP_019138373.1		
Corvus_brachyrhynchos	KFO55830.1		
Ovis_aries	XP_011961491.1		
Ovis_aries	XP_011961487.1		
Ovis_aries	XP_011961488.1	*Ovis_aries*	XP_011961489.1
Ovis_aries	XP_011961492.1		
Ovis_aries	NP_001124407.1		
Ovis_aries	NP_001124407.1		

Table 14 - Problematic aligned sequences of Algae, which were identified and manually removed, from the Phylogenetic tree.

Algae Regucalcin	
Species	**Blast results**
Fistulifera solaris	90%
Guillardia theta	24%

Table 15 - Problematic aligned sequences of Animals, which were identified and manually removed, from the Phylogenetic tree.

Animals Regucalcin		
Species	Blast results	Protein Accession Number
Pogona vitticeps	100% EXCLUDE	XP_020652482.1#
Manacus vitellinus	99% EXCLUDE	XP_008930214.2#
Tupaia chinensis	99% EXCLUDE	ELV14210.1#
Myotis brandtii	(96%, insertion) EXCLUDE	XP_014398972.1#
Nilaparvata lugens	44%	
Folsomia candida	(98%, insertion) EXCLUDE	XP_021956083.1#
Ailuropoda melanoleuca	99% EXCLUDE	EFB16351.1#
Cricetulus griseus	(99%;96% low quality) EXCLUDE	XP_007633712.1; XP_007629332.1#
Phascolarctos cinereus	99% EXCLUDE	XP_020848583.1#
Charadrius vociferus	99% EXCLUDE	XP_009880409.1#
Opisthocomus hoazin	(99%;100%) EXCLUDE	XP_009929456.1; XP_009929455.1#
Calypte anna	100% EXCLUDE	XP_008496314.1#
Lonchura striata domestica	98% EXCLUDE	XP_021402306.1
Peromyscus maniculatus bairdii	100% EXCLUDE	XP_015844818.1; XP_015844821.1; XP_015844822.1#
Cyprinus carpio	(91%, insertion) EXCLUDE	KTF96356.1#
Pygocentrus nattereri	(92%, insertion) EXCLUDE	XP_017557673.1#
Branchiostoma floridae	(86%, 76%)	
Branchiostoma belcheri	76%	
Strongylocentrotus purpuratus	99% EXCLUDE	XP_011676044.1#
Nilaparvata lugens	99% EXCLUDE	XP_022204098.1#
Acanthaster planci	99% EXCLUDE	XP_022089748.1#
Agrilus planipennis	(96%;97%)	
Ixodes scapularis	56%	
Papilio xuthus	99% EXCLUDE	XP_013166227.1#
Aethina tumida	80%	
Varroa destructor	100% EXCLUDE	XP_022650147.1#
Folsomia candida	(various blast) EXCLUDE	OXA48477.1; OXA46847.1; XP_021965762.1; OXA40016.1#
Zootermopsis nevadensis	(94%;74%) EXCLUDE	XP_021924408.1
Octopus bimaculoides	79%	
Crassostrea virginica	87%	

Crassostrea gigas	(93% insertion; 97%) EXCLUDE	EKC25142.1#
Diachasma alloeum	(100%;100%) EXCLUDE	XP_015116164.1; XP_015110079.1#
Fopius arisanus	100% EXCLUDE	XP_011306302.1#
Microplitis demolitor	(100%;73%) EXCLUDE	XP_008547468.1; XP_008547469.1#
Bombyx mori	100% EXCLUDE	XP_021206751.1; XP_004930716.1#
Papilio xuthus	99% EXCLUDE	XP_013166225.1#
Anoplophora glabripennis	90% EXCLUDE	
Drosophila pseudoobscura	99% EXCLUDE	KRS99979.1#
Drosophila bipectinata	83% EXCLUDE	
Dendroctonus ponderosae	99% EXCLUDE	ENN81654.1#
Nothobranchius furzeri	91% EXCLUDE	XP_015807699.1#
Oreochromis niloticus	99% EXCLUDE	XP_005466161.1#
Eufriesea mexicana	(92%, insertion) EXCLUDE	OAD58557.1#
Lottia gigantea	94%	
Mizuhopecten yessoensis	86%	
Trachymyrmex cornetzi	100% EXCLUDE	XP_018374114.1#
Harpegnathos saltator	100% EXCLUDE	XP_011153703.1#
Ooceraea biroi	99% EXCLUDE	XP_011332720.1#
Vollenhovia emeryi	99% EXCLUDE	XP_011881380.1#
Neodiprion lecontei	75%	
Aethina tumida	77%	
Rhagoletis zephyria	96%	
Bactrocera dorsalis	79%	
Bactrocera oleae	100% EXCLUDE	XP_014103595.1#
Aedes albopictus	(98%;97%) EXCLUDE	XP_019529541.1#
Halyomorpha halys	(99%;53%;88%) EXCLUDE	XP_014289311.1#
Aedes aegypti	100% EXCLUDE	EAT34066.1#
Aedes albopictus	100% EXCLUDE	XP_019534284.1#
Macrostomum lignano	(99%;85%) EXCLUDE	PAA52790.1#
Drosophila simulans	99% EXCLUDE	KMZ03555.1#
Bombyx mori	100% EXCLUDE	XP_021206751.1#
Agrilus planipennis	(97% writen isoforms) EXCLUDE	XP_018333452.1; XP_018333449.1#
Folsomia candida	98% EXCLUDE	XP_021965762.1#
Crassostrea gigas	97% EXCLUDE	XP_019926340.1#
Phantolops hodgsonii	EXCLUDE - contamination	XP_005955026.1
Phantolops hodgsonii	EXCLUDE - contamination	XP_005963615.1
Phantolops hodgsonii	EXCLUDE - contamination	XP_005958911.1
Ophisthorchis viverrini	EXCLUDE – poor annotation	OON18004.1
Cassostrea gigas	EXCLUDE	XP011448966.1

Lottia gigantea	EXCLUDE	ESP05381.1
Folsomia candida	EXCLUDE – False	OXA60774.1
Galendromus occidentalis	EXCLUDE – bacterial contamination	XP003738682.2
Hyalella azteca	EXCLUSE – False; contamination	XP_018028430.1

Table 16 - Problematic aligned sequences of Fungi, which were identified and manually removed, from the Phylogenetic tree.

Fungi Regucalcin		
Species	**Blast results**	**Protein Accession Number**
Cryptococcus neoformans var. grubii	(100%;99%;99%)	OXC69596.1; OXH29903.1; OWZ70398.1#
Penicillium brasilianum	99% EXCLUDE	CEJ57896.1#
Emmonsia crescens	97%	
Emmonsia parva	90%	
Hortaea werneckii EXF-2000	90%	
Alternaria alternata	100% EXCLUDE	OWY41493.1#
Fusarium fujikuroi	(99%;99%;99% CCT66319.1 different name and region) EXCLUDE	KLP21447.1;KLO89672.1#
Kluyveromyces marxianus	100% EXCLUDE	BAP73275.1#
Alternaria alternata	99% EXCLUDE	OAG14347.1#
Drechmeria coniospora	99% EXCLUDE	ODA78235.1#
Cryptococcus neoformans var. grubii	(98%, different region)	

Table 17 - The isoforms identified and removed in the Regucalcin dataset (C).

Species	Protein accession number
Oikopleura dioica	CBY20807.1
Nilaparvata lugens	XP_022203265.1
Trichomalopsis sarcophagae	OXU28709.1
Bactrocera latifrons	XP_018802379.1
Bemisia tabaci	XP_018902740.1
Branchiostoma floridae	EEN58249.1
Gallus_gallus	XP_015156358.1
Nilaparvata lugens	XP_022187325.1
Peromyscus maniculatus bairdii	XP_015844817.1
Pogona vitticeps	XP_020652480.1
Tupaia_chinensis	XP_006164595.1

2. Regucalcin group phylogenies with duplications

2.1. Non-Bilateria

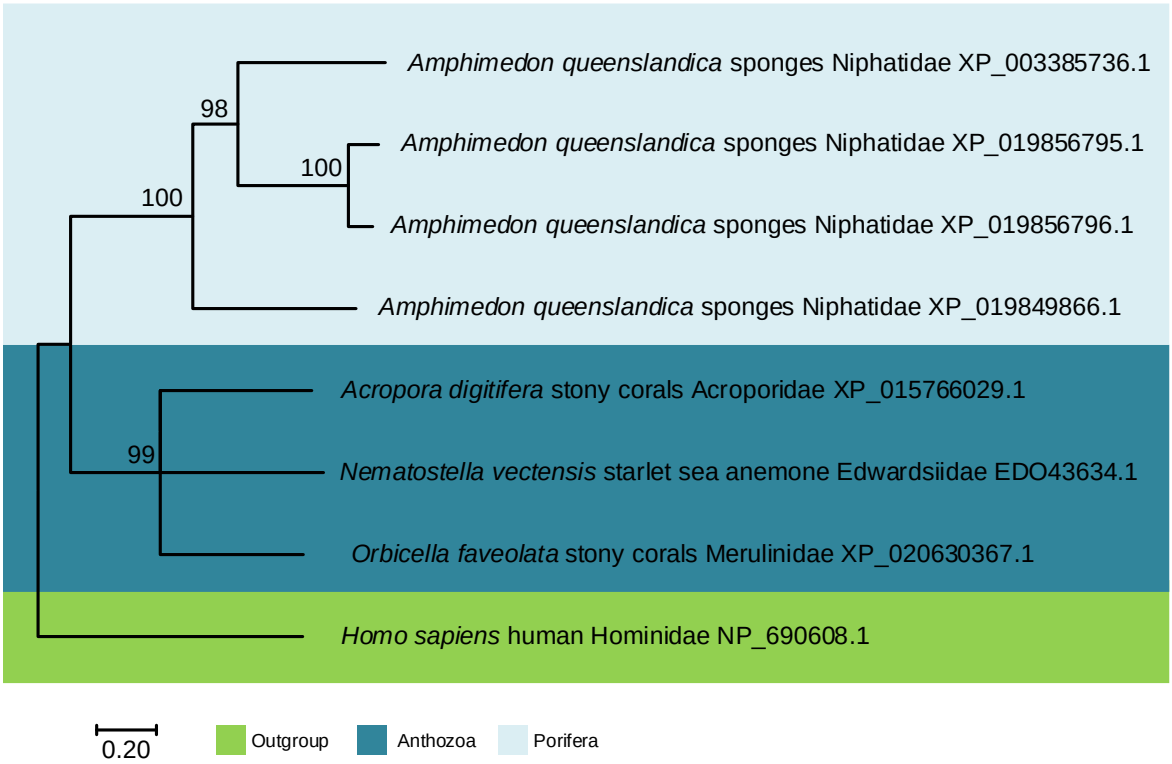


Figure S 2 - Non-Bilateria *Regucalcin* Bayesian phylogeny. Notice that the *Homo sapiens* sequence is just present, as an outgroup.

2.2. Protostomia Lophotrochozoa

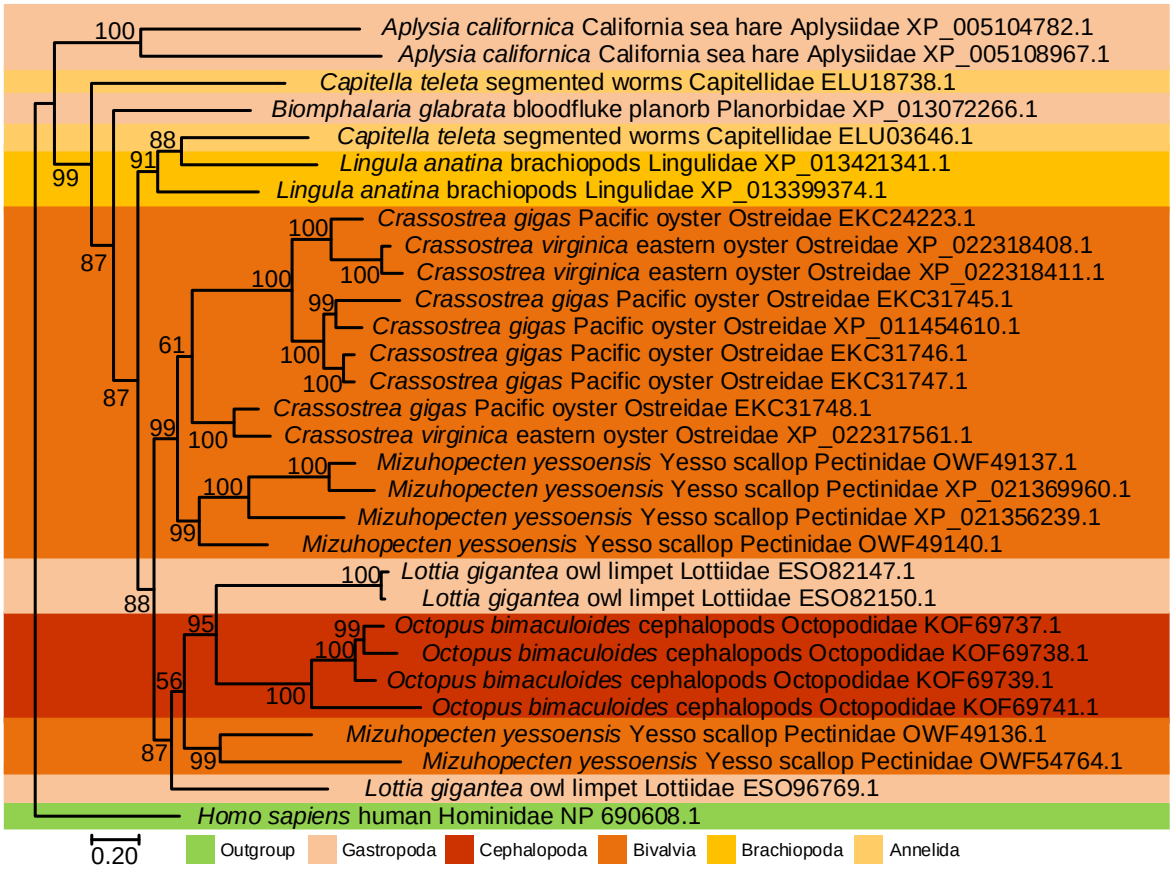


Figure S 3 - Protostomia-Lophotrochozoa *Regucalcin* Bayesian phylogeny. Notice that the *Homo sapiens* sequence is just present, as an outgroup.

2.3. Insecta – Hemiptera/Blattodea

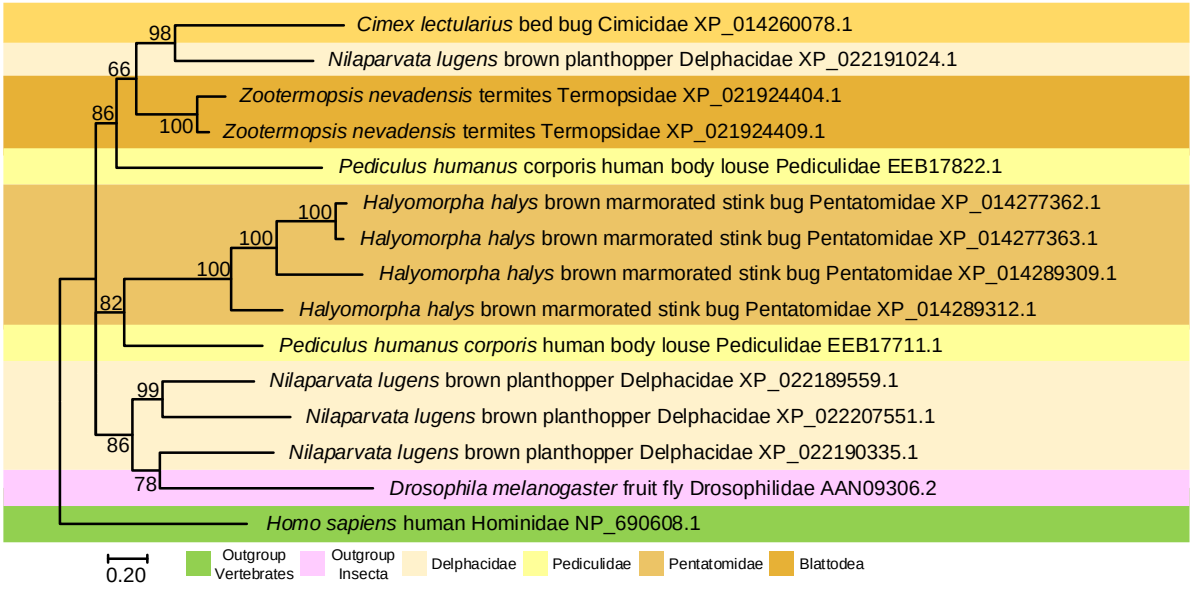


Figure S 4 - Hemiptera *Regucalcin* Bayesian phylogeny. Notice that the *Homo sapiens* (representing vertebrates) *and Drosophila melanogaster* (representing the remaining insects) sequences are just present as outgroups, represented in green and in pink, respectively.

2.4. Insecta – Coleoptera

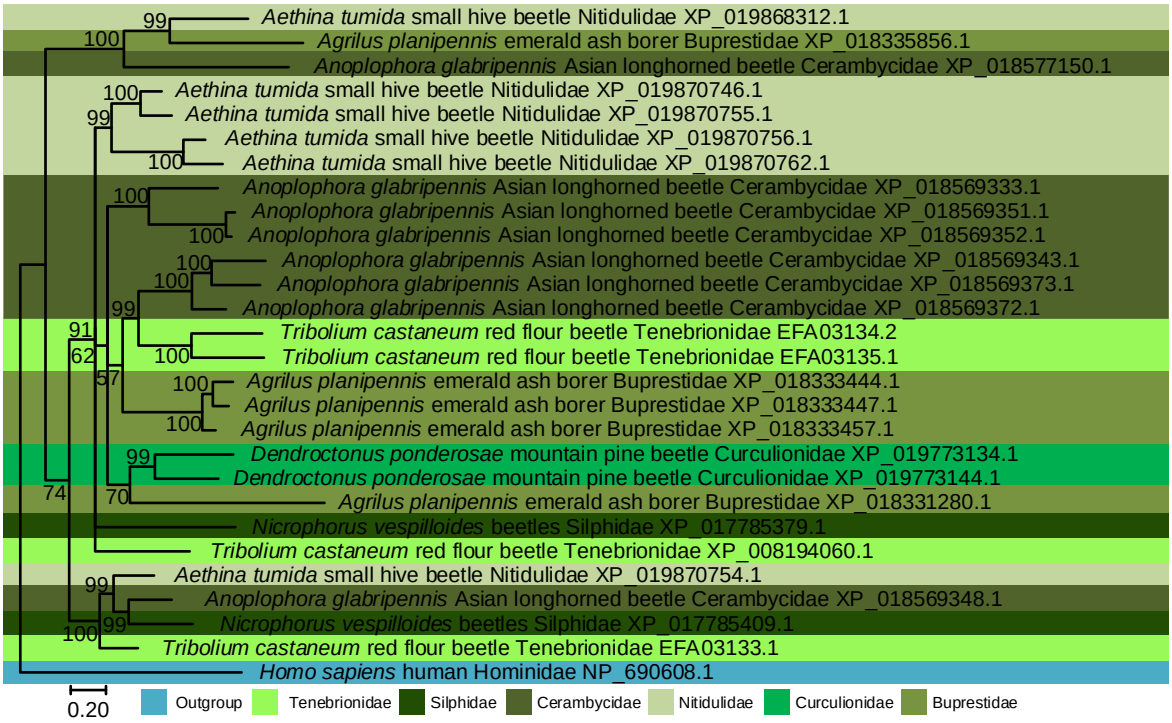


Figure S 5 - Coleoptera *Regucalcin* Bayesian phylogeny. Notice that the *Homo sapiens* sequence is just present as outgroup, represented in blue.

2.5. Insecta – Hymenoptera

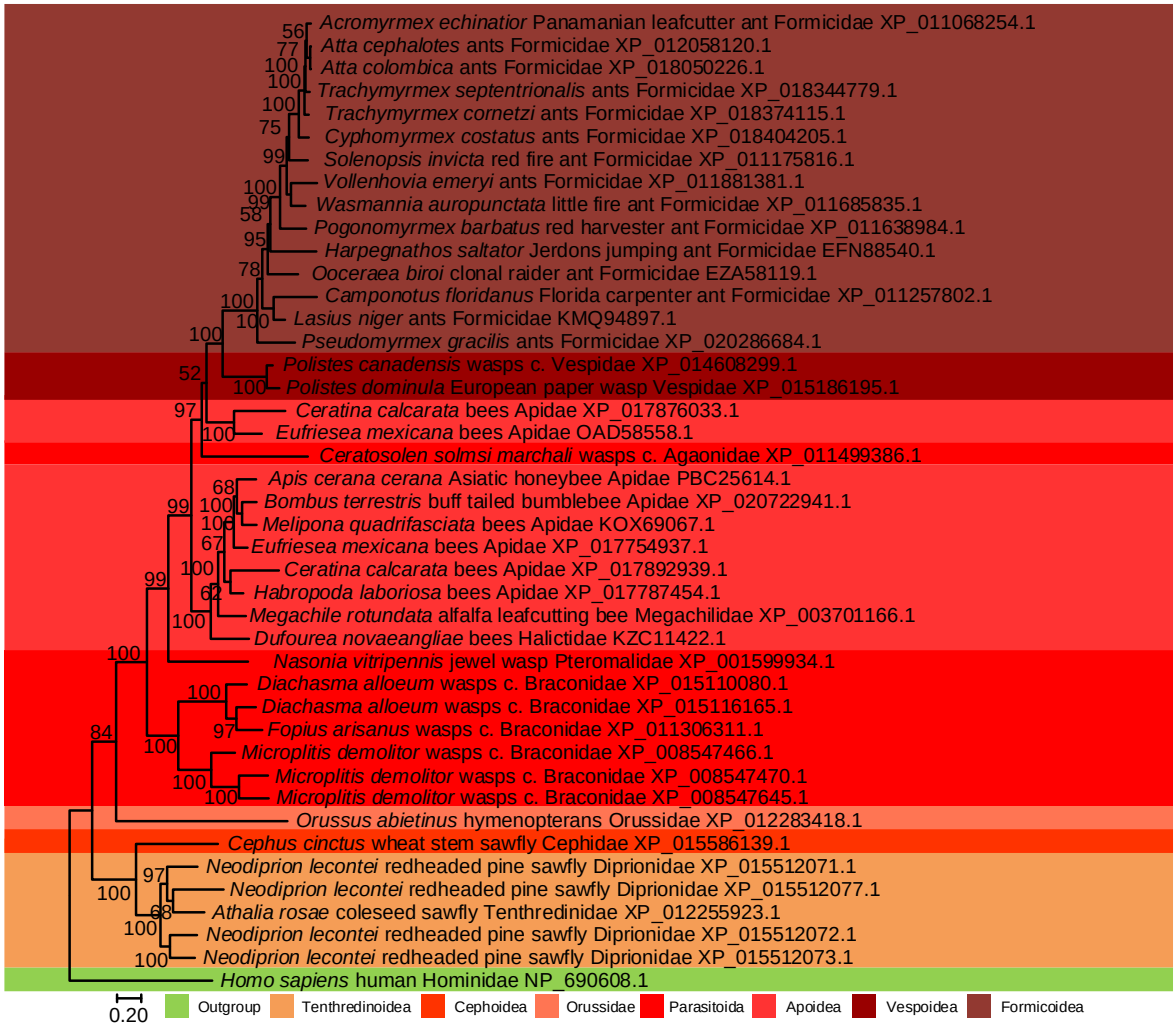
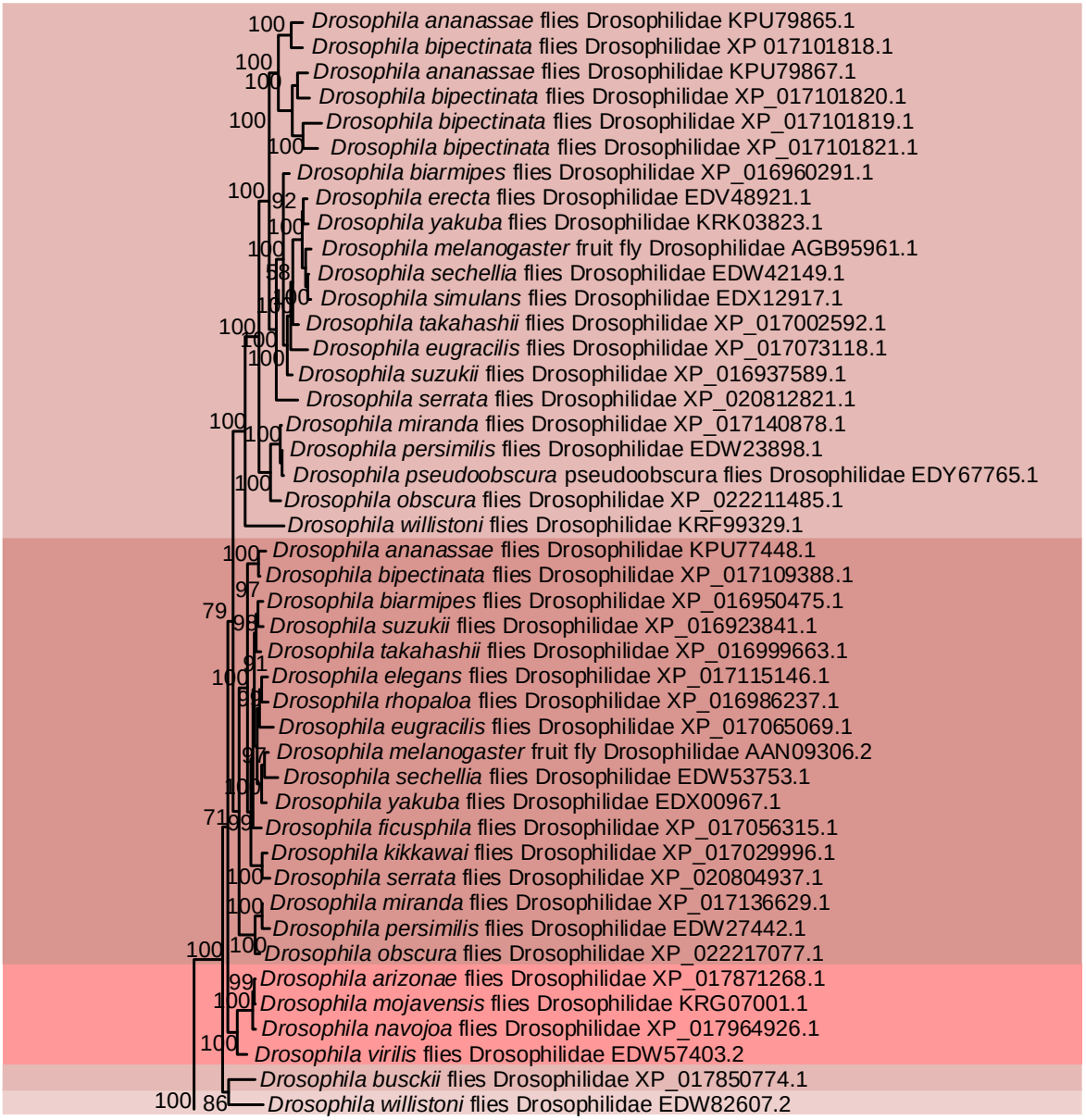


Figure S 6 - Hymenoptera *Regucalcin* Bayesian phylogeny. Notice that the *Homo sapiens* sequence is just present as outgroup, represented in green.

2.6. Insecta – Diptera



(The image continues on the next page)

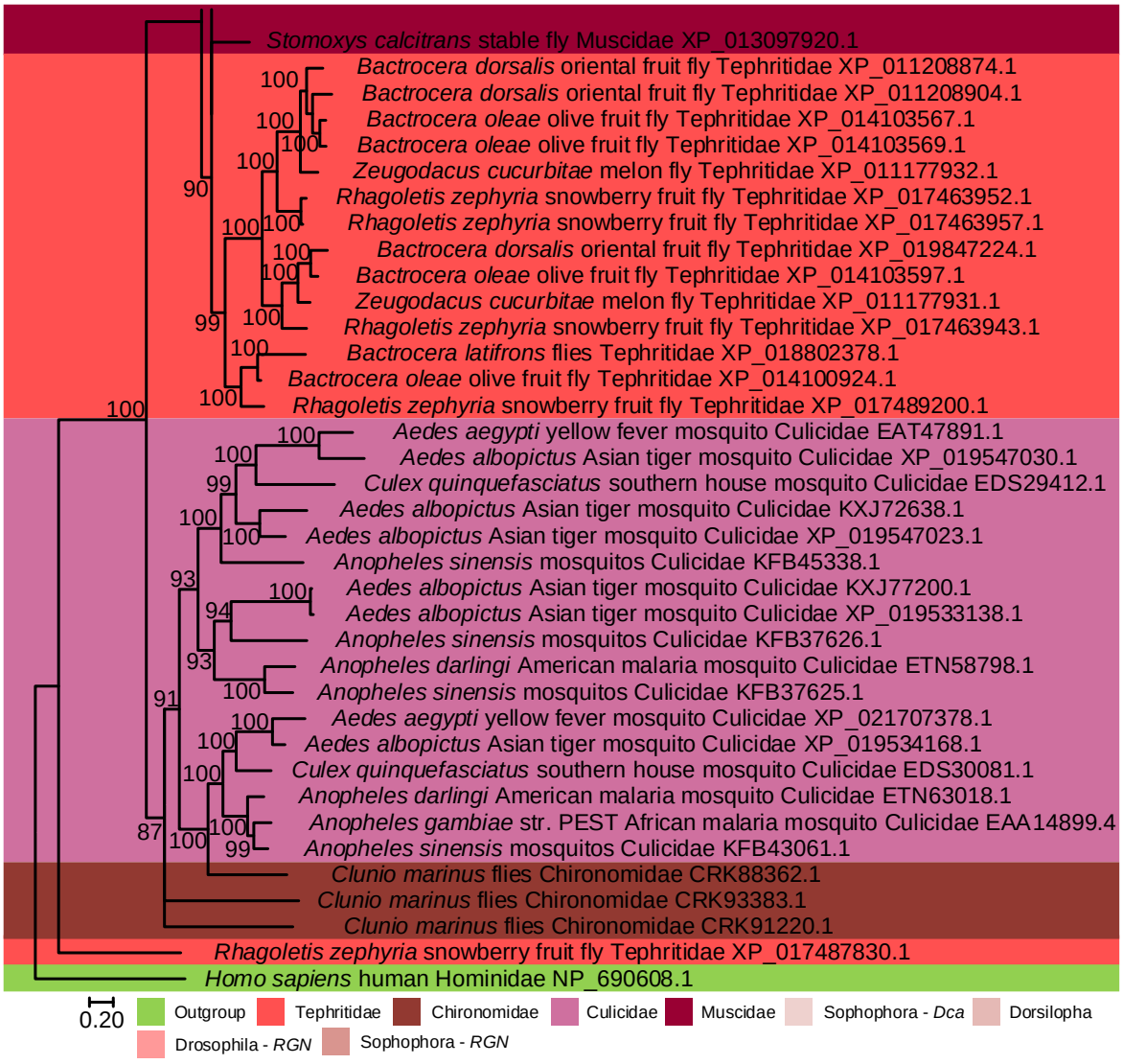


Figure S 7 - Diptera *Regucalcin* Bayesian phylogeny. Notice that the *Homo sapiens* sequence is just present as outgroups, represented in green.

2.7. Insecta – Lepidoptera

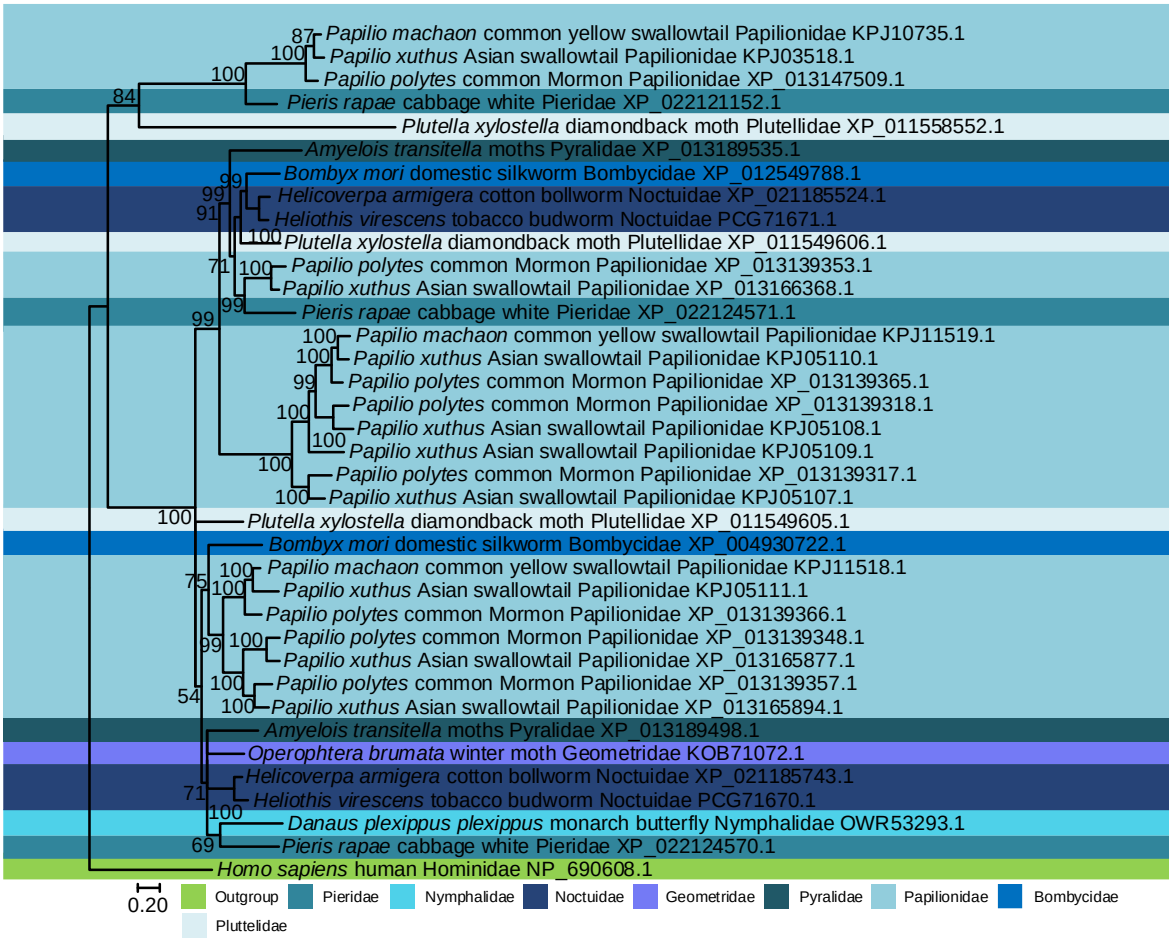


Figure S 8 - Lepidoptera *Regucalcin* Bayesian phylogeny. Notice that the *Homo sapiens* sequence is just present as outgroups, represented in green.

2.8. Protostomia Non-Lophotrochozoa

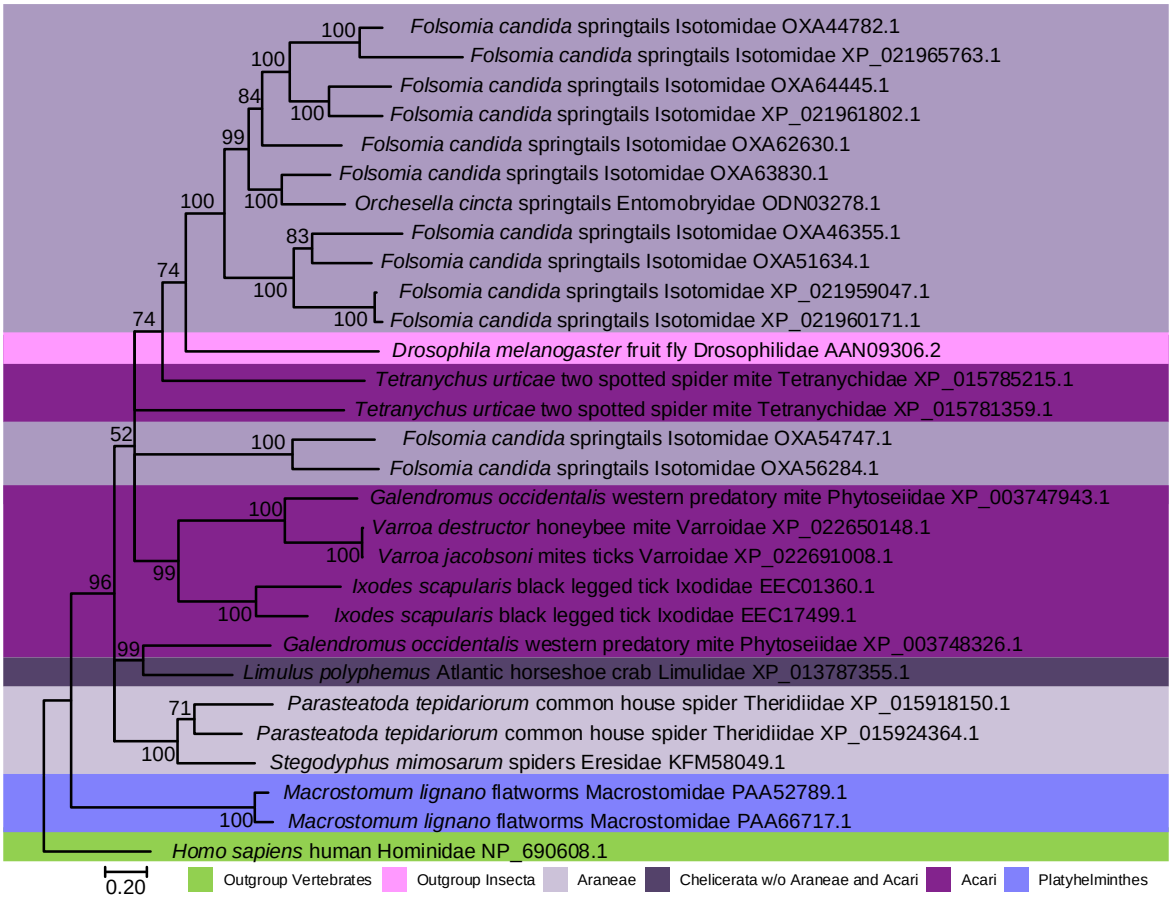


Figure S 9 - Protostomia Non-Lophotrochozoa *Regucalcin* Bayesian phylogeny. The colours pink and green represent both insect and vertebrates outgroups, respectively.

2.9. Basal Deuterostomians

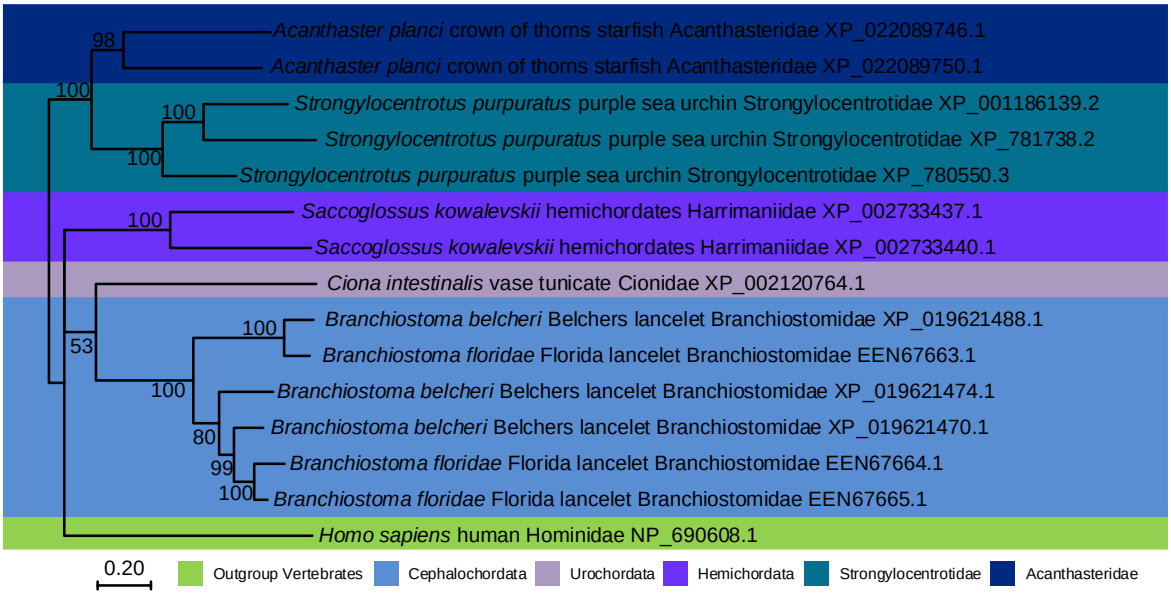
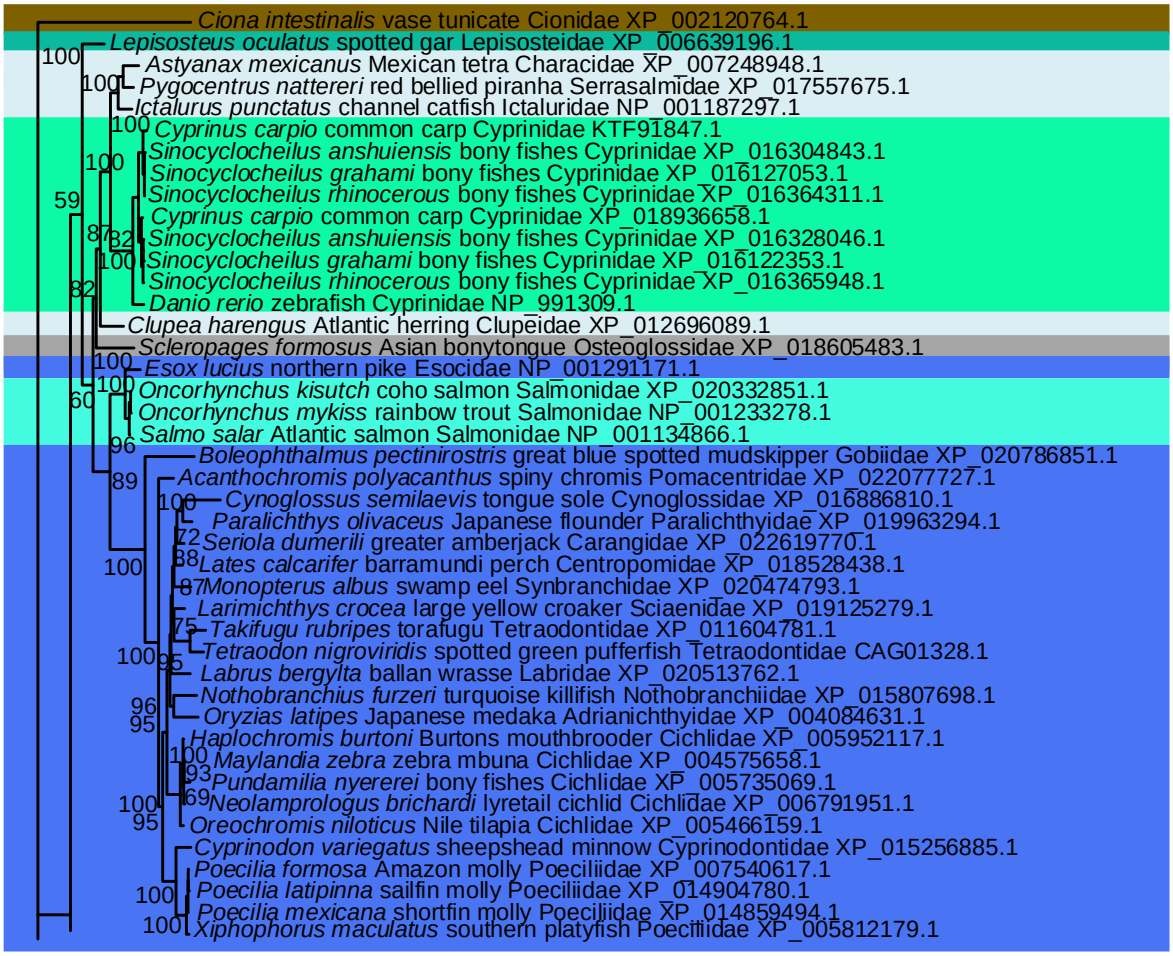


Figure S 10 - Basal Deuterostomia *Regucalcin* Bayesian phylogeny. Notice that the *Homo sapiens* sequence is just present, as an outgroup.

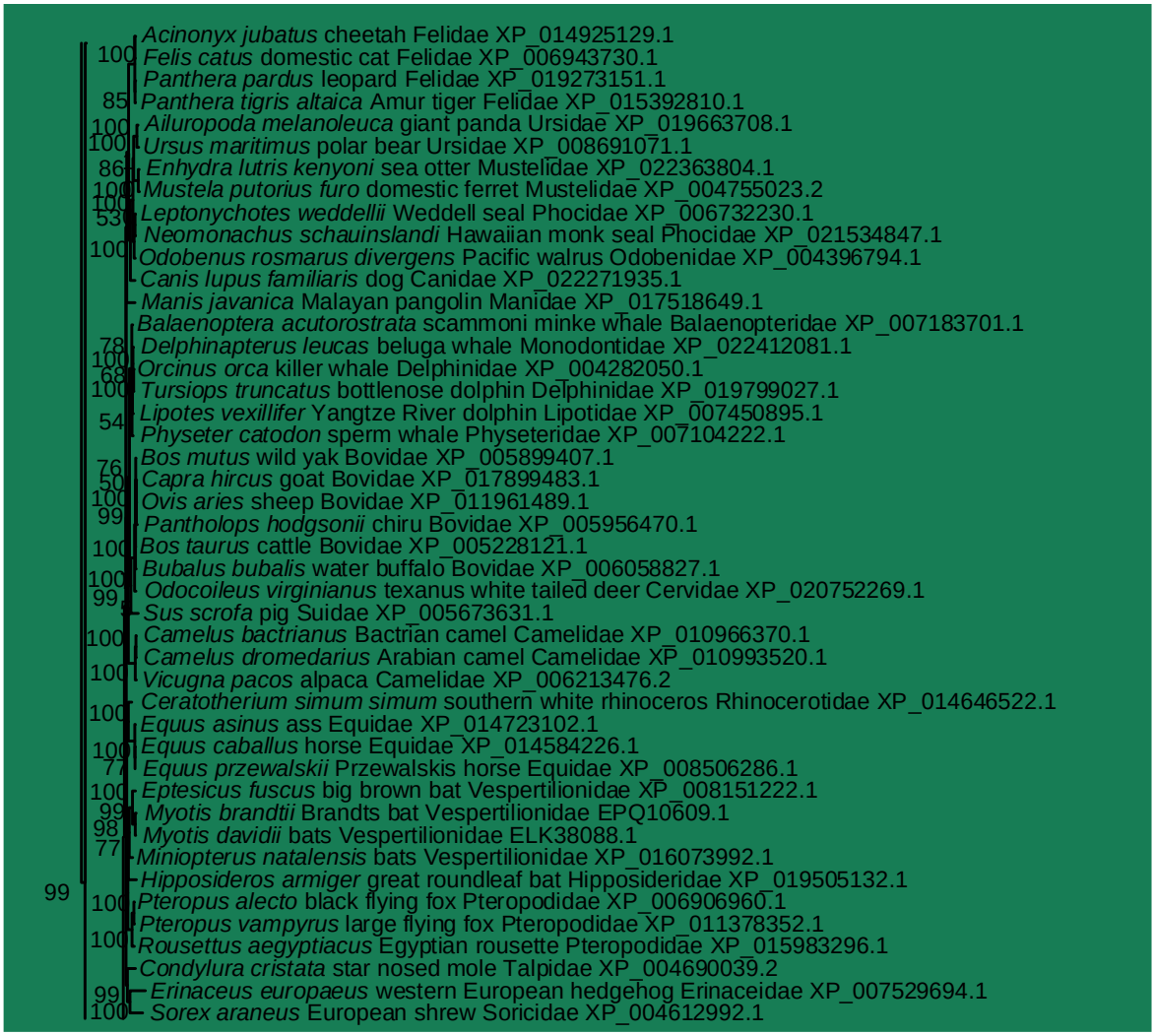
2.10. Vertebrates



(The image continues on the next page)

(The image continues on the next page)

(The image continues on the next page)



(The image continues on the next page)

Figure S 11 - Vertebrate *Regucalcin* Bayesian phylogeny. Notice that the *Ciona intestinalis* sequence is just present, as an outgroup, represented in olive green.

3. Interaction proteins

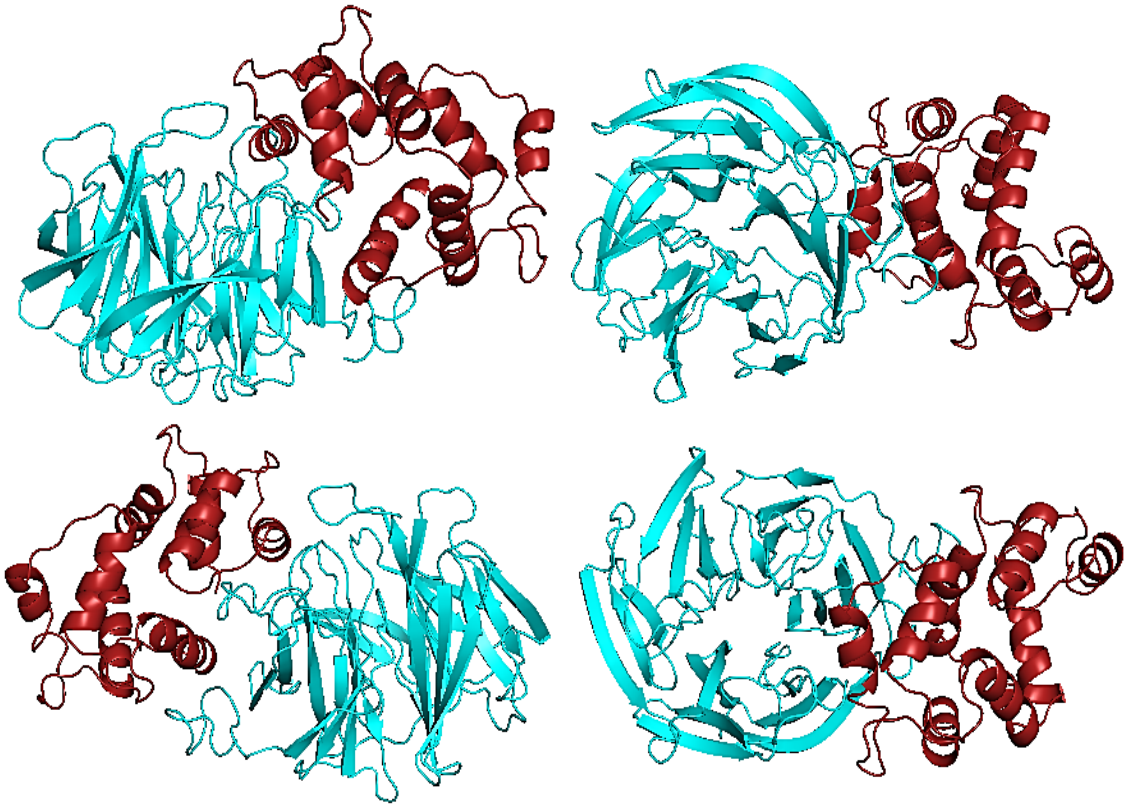


Figure 36 – Regucalcin interaction with calcium interacting Calmodulin protein.

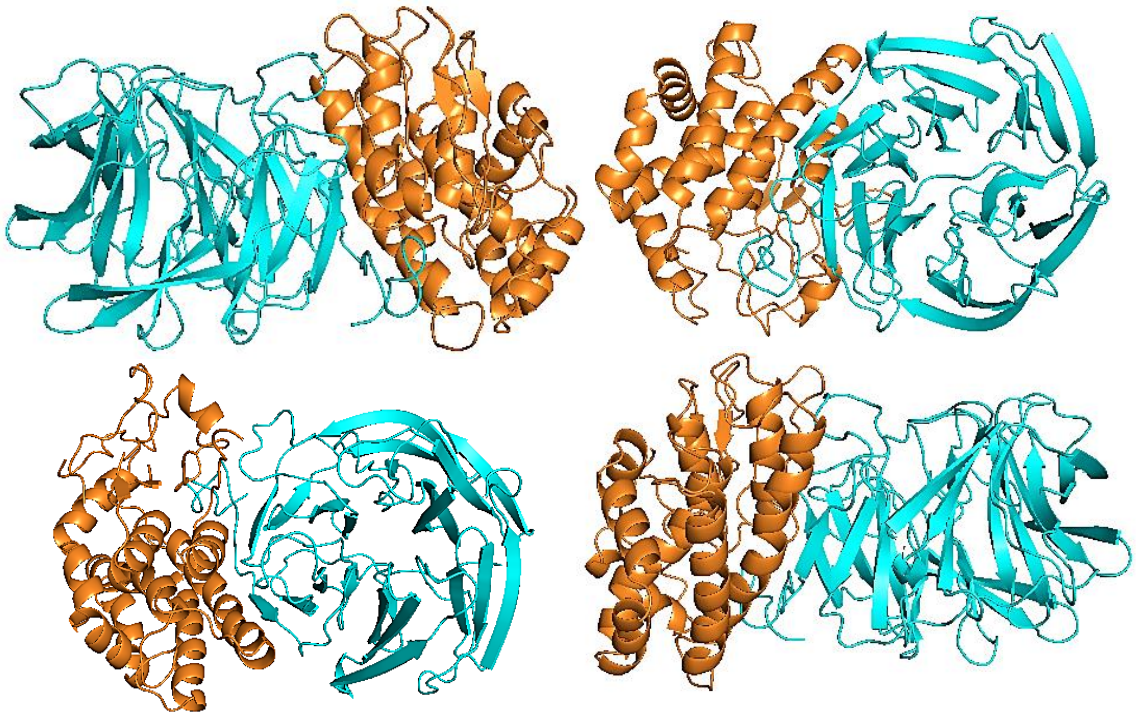


Figure 37 - Regucalcin interaction with the Ascorbic Acid synthesis interacting Glutathione S Transferase protein.