Evolucija ključnih peroksisomalnih proteina

Mišetić, Hrvoje

University of Zagreb

Faculty of Science

Department of Biology

Hrvoje Mišetić

Evolution of core peroxisomal proteins

Evolucija ključnih peroksisomalnih proteina

Master Thesis

Zagreb, 2019.

This master thesis was conducted at the CRG-Centre for Genomic Regulation, Barcelona, Spain, under the supervision of research professor Toni Gabaldón, PhD, and co-supervision of Assoc. Prof. Damjan Franjević, PhD. The thesis was submitted for evaluation to the Department of Biology at the Faculty of Science, University of Zagreb with the aim of obtaining the title Master in Molecular Biology.

BASIC DOCUMENTATION CARD

University of Zagreb

Faculty of Science

Department of Biology

Master thesis

Evolution of core peroxisomal proteins

Hrvoje Mišetić

Roosveltov trg 6, 10000 Zagreb, Hrvatska

There are a lot of unanswered questions about the evolutionary origin of the peroxisome. About ten years ago a phylogenetic analysis of peroxisomal proteome was conducted during which the minimal ancestral eukaryotic peroxisomal proteome was discovered. These days the larger amount of genomic data provides an ideal framework to analyze the origin of the ancestral peroxisome. To address this question, phylogenetic trees for each protein from the proteome were reconstructed and complemented with predictions of subcellular localization and protein domain information to perform a detailed phylogenetic analysis. Peroxins Pex1, Pex2, Pex4 and Pex10 show common origin with the components of ERAD, Pex5 with components of anaphase promoting complex/cyclosome (APC/C) while Pex14 doesn't have evolutionary relationship with any molecular system. Pxa1 and Pxa2 that are involved in lipid transport show independent origin from its mitochondrial and cell membrane counterparts. Adding members from newly sequenced eukaryotic phyla and Asgard group didn't change previously known phylogeny of catalase. Fox1 and Faa2 that are involved with peroxisomal β-oxidation show independent origin from the same process in mitochondria while Fox2 that catalyzes second step of β-oxidation is probably derived from mitochondria.

(80 pages, 32 figures, 11 tables, 99 references, original in: English)

Thesis deposited in the Central Biological Library

Keywords: peroxisome, evolution, phylogenomics

Supervisor/Co-supervisior: research professor Toni Gabaldón, PhD / Assoc. Prof. Damjan Franjević, PhD

Reviewers: Assoc. Prof. Damjan Franjević, PhD; Prof. Biljana Balen, PhD; Asst. Prof. Silvija Černi, PhD

Thesis accepted: January 31st, 2019

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek Diplomski rad

Evolucija ključnih peroksisomalnih proteina

Hrvoje Mišetić

Roosveltov trg 6, 10000 Zagreb, Hrvatska

Postoji niz neodgovorenih pitanja o evolucijskom podrijetlu peroksisoma. Prije desetak godina provedena je filogenetska analiza peroksisomalnog proteoma u kojoj je otkriven minimalni ancestralni peroksisomalni proteom. Današnja velika količina genomskih podataka pruža idealne uvjete za analizu podrijetla ancestralnog peroksisomalnog proteoma. Kako bi se riješilo to pitanje, filogenetska stabla su napravljena za svaki protein u proteomu i nadopunjena su predikcijama sub-stanične lokalizacije i informacijama o proteinskim domenama s namjerom da se provede detaljna filogenetska analiza. Peroksini Pex1, Pex2, Pex4 i Pex10 imaju zajedničko podrijetlo s komponentama sustava ERAD, Pex5 s komponentama kompleksa koji promovira anafazu (engl. *anaphase promoting complex*) dok Pex14 nema evolucijsku povezanost s nijednim molekularnim sustavom. Pxa1 i Pxa2 koji imaju ulogu u transportu masnih kiselina imaju neovisno podrijetlo od odgovarajućih proteina u mitohondriju i staničnoj membrani. Dodavanje predstavnika nedavno sekvenciranih eukariotskih koljena i članova Asgard grupe ne mijenja postojeću filogeniju katalaze. Fox1 i Faa2 koji imaju ulogu u peroksisomalnoj β-oksidaciji masnih kiselina imaju neovisno podrijetlo od proteina koji sudjeluju u istom procesu smještenom u mitohondriju-dok Fox2 koji katalizira drugi korak oksidacije masnih kiselina ima podrijetlo iz mitohondrija.

(80 stranica, 32 slike, 11 tablica, 99 literaturnih navoda, jezik izvornika: engleski)

Rad je pohranjen u Središnjoj biološkoj knjižnici

Ključne riječi: peroksisom, evolucija, filogenomika

Voditelj/Suvoditelj: prof. dr. sc. Toni Gabaldón / izv. prof. dr. sc. Damjan Franjević

Ocjenitelji: izv. prof. dr. sc. Damjan Franjević, prof. dr. sc. Biljana Balen, doc. dr. sc. Silvija Černi

Rad prihvaćen: 31.siječanj, 2019.

1. INTRODUCTION 1

1.1. Discovery of peroxisome 1
1.2. Metabolic role 2
1.3. Common features of peroxisomes 5
1.4. Evolutionary origin of peroxisomes 7
1.5. Hypothesis and aim of study 9

2. MATERIALS & METHODS 10

2.1. Database and query sequences 10
2.2. BLAST (Basic Local Alignment Search Tool) search 11
2.3. Profile HMM (Hidden Markov Model) search 11
2.4. Sequence clustering 12
2.5. Reconstruction of phylogenetic trees 13
2.6. Subcellular localization prediction 13
2.7. Domain annotation 14
2.8. Tree visualization 14

3. RESULTS 15

3.1. BLAST results 15
3.2. HMM profile search 16
3.3. Clustering results 17
3.3.1. Bacteria 17
3.3.2. Archaea 21
3.4. Reconstruction of phylogenetic trees 23
3.4.1. Eukaryotic phylogenetic trees 25
3.4.2. Peroxisomal trees 37
3.4.3. Phylogenetic trees based on orthology search 43

4. DISCUSSION 56

4.1. Catalase 56
4.2. Lipid transport 57
4.3. Peroxisomal protein import 58
4.4. Peroxisomal β-oxidation 62

5. CONCLUSION 65

6. REFERENCES 66

7. SUPPLEMENTARY 71

8. CURRICULUM VITAE 73

1. INTRODUCTION

1.1. Discovery of peroxisome

Peroxisome is a cellular organelle bound by a single lipid membrane that contains variety of different enzymes depending on the species, specific tissue or the environmental conditions (Gabaldón, 2010). Firstly, peroxisomes were noted by Johannes Rhodin in 1954 while studying the morphology of proximal tubule cells from mouse kidney and they were initially named microbodies (Rhodin, 1954). Afterwards their presence in different mammalian tissues was noticed in various electron microscopy studies. A first biochemical characterization of microbodies was done by Christian de Duve and his colleagues in the early 1960s. They isolated organelles from the rat liver and separated microbodies from mitochondria, lysosomes and microsomes using miscellaneous types of density gradient centrifugation to study its biochemical features (De Duve and Baudhuin, 1966). In these organelles, they discovered several oxidases that oxidize their substrates while reducing oxygen to hydrogen peroxide (H_2O_2). Also, presence of two classes of enzymes capable of reducing H_2O_2, peroxidases and catalases, was noted. Because of the production and then eventual degradation of hydrogen peroxide, which is harmful to the cell, this newly characterized organelle was named peroxisome by Christian de Duve.

In the following years peroxisomes were biochemically characterized in detail mostly in mammalian, plant and fungal cells. Those characterizations revealed remarkable diversity not only across different species but within different tissues of multicellular organism or within unicellular eukaryotic organisms dependent on the environmental or developmental conditions. This organelles range in size between 0.1 and 1.5 μm in diameter and have dense matrices containing metabolic enzymes which can form structured and electron-dense crystalloid cores (Smith and Aitchison, 2013). Usually they have a spherical shape but their shape can change depending on the cell type and environment, so they can even be elongated or form reticula (Schrader and Fahimi, 2006). Besides that, peroxisomes can conditionally increase in size and number in coordination with morphological changes in other subcellular compartments (Jung et al., 2013). A single boundary lipid bilayer membrane distinguishes them from other microbodies found in eukaryotic microorganisms such as hydrogenosomes and mitosomes that are both related to mitocohondria (Muller et al., 2012).

Peroxisomes are present in representatives of all major eukaryotic lineages, which indicate that the organelle has arisen before LECA (Last Eukaryotic Common Ancestor). Even though its enzymatic content varies a lot across species, there are properties that are

common to all peroxisome such as set of proteins involved in peroxisome biogenesis and maintenance, which indicate a single evolutionary origin (Gabaldón, 2010).

1.2. Metabolic role

Peroxisomes harbor more than 50 different enzymes that are involved in various metabolic pathways including both anabolic and catabolic ones. Most common metabolic processes localized in the peroxisomes among wide range of species are oxidative reactions of uric acid, amino acids and fatty acids.

Fatty acid β-oxidation occurs in almost all peroxisomes and it is considered to be their original function (Gabaldón et al., 2006). It is important to note that in animal cells, fatty acids are oxidized in both peroxisome and mitochondria, while in yeasts and plants they are exclusively oxidized in the peroxisome. The major difference between the two systems is that in mitochondrial pathway first step is catalyzed by a FAD-dependent acyl-CoA dehydrogenase. The $FADH_2$ passes its electrons into the respiratory chain. First step of peroxisomal pathway is catalyzed by a FAD-dependent acyl-CoA oxidase. Here the electrons from $FADH_2$ are directly transferred to O_2,which results in the production of H_2O_2 which is a hallmark molecule of the peroxisome (Gabaldón, Ginger and Michels, 2016). The oxidative branch of the pentose-phosphate pathway, that is usually localized in the cytosol, appears in the peroxisomes in plants and parasitic protozoa (Kruger & von Schaewen, 2003).

Besides the oxidative reactions, peroxisomes are involved in lipid biosynthesis. In animal cells cholesterol and dolichol are synthesized both in peroxisome and endoplasmic reticulum. Bile acids, which are derived from cholesterol, are partially synthesized in the peroxisomes of the liver. Also, peroxisome has enzymes that are involved in the synthesis of plasmalogens, which are phospholipids that are important membrane components in some tissues, especially heart and brain (Cooper and Hausman, 2000). Biosynthesis of ether-lipids occurs in peroxisomes because the enzymes dihydroxyacetonophosphate acyltransferase and alkyldihydroxyacetonephosphate synthase both reside in peroxisome and are necessary for the formation of the ether bond (Hajra, 1995).

In plants, peroxisomes play two significant roles. First one is the conversion of stored fatty acids into carbohydrates via glyoxylate cycle which is a variant of citric acid cycle. In this cycle isocitrate, instead of being degraded to CO_2 and α-ketoglutarate, is converted to succinate and glyoxylate that reacts with a molecule of acetyl CoA to yield malate, which is converted to oxaloacetate and used for glucose synthesis (Figure 1.). This is one of the crucial biochemical processes for providing energy and raw materials during the growth of the germinating plant (Cooper and Hausman, 2000). Because of this process organelles in the germinating plant seedlings were called glyoxysomes even though they are equilibrated

at the same density in sucrose gradient as the peroxisomes from other species and also contain catalase, glycolate oxidase and urate oxidase (Breidenbach, Kahn and Beevers, 1968). Later, it became clear that glyoxysomes are subset of peroxisomes, but their initial name is continued to be widely used.

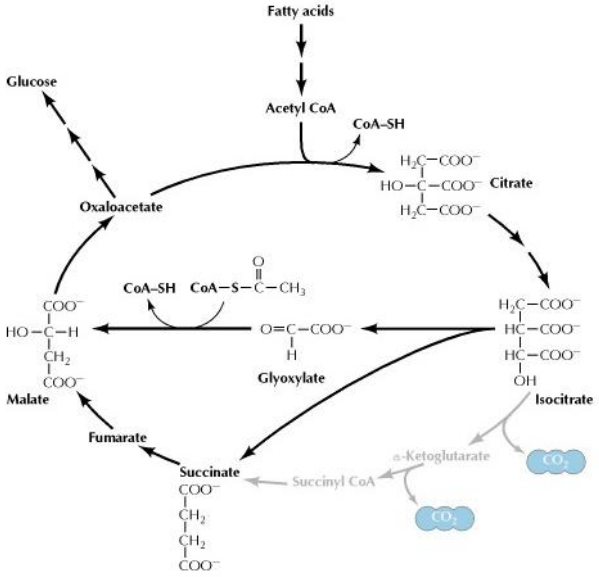


Figure 1. Glyoxylate cycle – conversion of acetyl-CoA to succinate for the synthesis of carbohydrates (Cooper and Hausman, 2000).

Second significant role in which peroxisomes from leaves are involved together with mitochondria and chloroplast is photorespiration.Photorespiration is a process which metabolizes a side product formed during photosynthesis. Usually during the Calvin cycle, CO_2 is added to the five-carbon sugar ribulose-1,5-bisphosphate but sometimes O_2 is added instead, so the final product of the Calvin cycle is 3-phosphoglycerate and phosphoglycolate instead of two molecules of 3-phosphoglycerate. Because the phosphoglycolate is not a useful product, it is converted to glycolate that is then transported to peroxisomes, where it is oxidized and converted to glycine which is transported to mitochondria and converted to serine. The serine is moved back to peroxisome to be converted to gycerate that can then be used in choroplasts to re-enter the Calvin cycle (Figure 2.). As seen, peroxisome plays a major role in allowing most of the carbon in glycolate to be recovered and used (Cooper and Hausman, 2000).

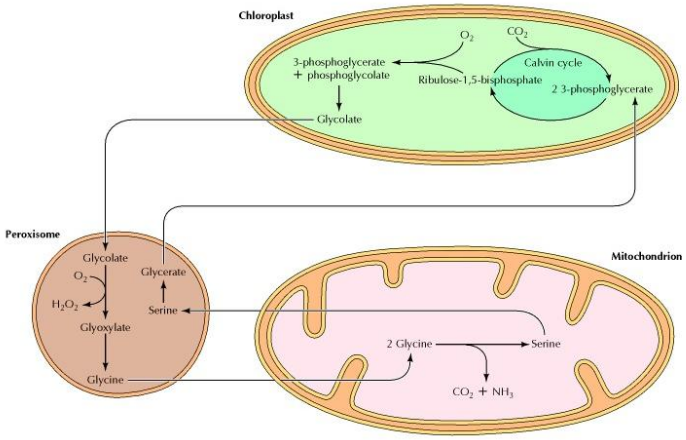


Figure 2. Photorespiration-the metabolism of phosphoglycolate that is formed by adding O_2 to the ribulose-1,5-bisphosphate (Cooper and Hausman, 2000).

Besides plants, fungi also show remarkable variety of specific metabolic processes that reside inside of peroxisome. Synthesis of biotin and a span of secondary metabolites suchlike antibiotics and toxins like polyketides are located in peroxisome (van der Klei and Veenhuis, 2013). Furthermore, a type of peroxisome called Woronin body which has a role in maintaining cellular integrity by plugging septal pores to stop cytoplasmic bleeding of damaged hyphae has been noticed in some ascomycetes (Jedd and Chua, 2010). Similarly, some glycolytic enzymes that usually reside in cytosol can be found in the peroxisomes of many fungi. The reason for that is the presence of cryptic targeting signals that are revealed after alternative splicing or stop-codon read-through. Glycolytic enzymes inside of fungal peroxisomes are involved in maintaining redox and ATP/ADP ratio homeostasis (Freitag, Ast and Bölker, 2012).

For the protists of the groups Kinetoplastea and Diplonemida, enzymes of glycolytic pathway and gluconeogenic pathway are present in the subtype of peroxisomes called glycosomes. There are two types of glycosomes that are found in the organisms that have this organelles. First type of glycosomes are lyoglycosomes which are found free in the cytosol of the cell where they form chains and are more abundant in healthy cells. They tend to be fewer electrons dense. While the other type, desmoglycosomes, are not free in the cell but attached to the other organelles such as myofibrils, mitochondria and endoplasmic reticulum. They do not form groups but stay separate and are associated to high amount of proteins which results in high electron density (Rybicka, 1996).

Additionally to everything mentioned, peroxisomes play a significant role in a series of cellular signaling processes. Most important to point out are their complex interactions with lipid droplets and mitochondria and generation of ether-lipids which are a major factor in cellular signaling. Peroxisomal signaling mechanisms are still much unknown but suspected

to be very relevant and influential to many disorders including cancer, obesity-related diabetes and degenerative neurologic diseases (Lodhi and Semenkovich, 2014).

From everything mentioned, it is clear that enzymatic content of peroxisomes is quite diverse among species to the extent that they were firstly classified as different organelles which denote high level of evolutionary plasticity (Gabaldón, 2010).

1.3. Common features of peroxisomes

Even though peroxisomes carry out a wide variety of metabolic processes across different taxonomic lineages, there are two main features that are common to all of them and those are sets of proteins involved in peroxisome biogenesis and maintenance. The biogenesis and maintenance processes comprise proteins involved in protein import and organelle division that are present in all types of peroxisomes (Gabaldón, 2010).

Contrary to mitochondria and chloroplast, peroxisomes do not have an organelle genome. All peroxisomal proteins are encoded by nuclear genome and translated by ribosomes. These proteins need to be imported into the peroxisome in a certain way. Matrix proteins and membrane proteins are the two types of proteins that are incorporated into the peroxisome, but both of them have distinct targeting signals and mechanisms of import (Smith and Aitchison, 2013) and only the matrix protein import is widely conserved so it will be described in more detail.

Matrix proteins are post-translationally routed from the cytosol to peroxisomes by peroxisomal targeting signals (PTSs). There are two main signals: PTS1 and PTS2. PTS1 is a tripeptide with a sequence (Ser/Ala/Cys)(Lys/Arg/His)(Leu/Met/Ile) located on the C-terminus of the protein and it is found in most matrix proteins. PTS2 is bipartite signal with a consensus sequence (Arg/Lys)(Leu/Val/Ile)$(X)_5$(His/Cin)(Leu/Arg), which is located near the amino terminus and it is much more less often that the PTS1 motif (Smith and Aitchison, 2013). Those targeting signals are recognized by the molecular machinery that takes up the matrix protein inside of the peroxisome. Molecular machinery is made of peroxins which are proteins encoded by PEX genes that are involved in peroxisome biogenesis. In the first step of the peroxisomal protein import, peroxisomal protein is recognized by the import receptor peroxin 5 (Pex5) and the receptor-cargo complex is routed to the peroxisomal docking complex made of Pex13, Pex14 and Pex5. After the receptor-cargo complex enters into the peroxisome, import receptor can be recycled and moved back to the cytosol ready for the new import cycle or it can be disposed in case it is dysfunctional through a quality control pathway. In the quality control pathway, Pex5 from the peroxisome matrix becomes an integral membrane protein that is polyubiquitylated by a ubiquitylation cascade. Ubiquitylation cascade is made of E2 ubiquitin-conjugating enzyme (Ubc) 4 and Ubc5, Pex2 and Pex10 E3

ligases that are connected to the docking complex via Pex8. The polyubiquitylated Pex5 is then released from the membrane by the action of Pex1 and Pex6 that both contain AAA-type ATPase domains and then directed to the proteosomal degradation (Figure 3.a). In the alternative cycling pathway, Pex5 is monoubiquitylated by an ubiquitylation machinery made out of the E2 Pex4 and the E3 Pex12. The monoubiquitylated receptor is released in the cytosol while the ubiquitin moiety is removed and the receptor is available for another cycle of import (Figure 3.b) (Smith and Aitchison, 2013).

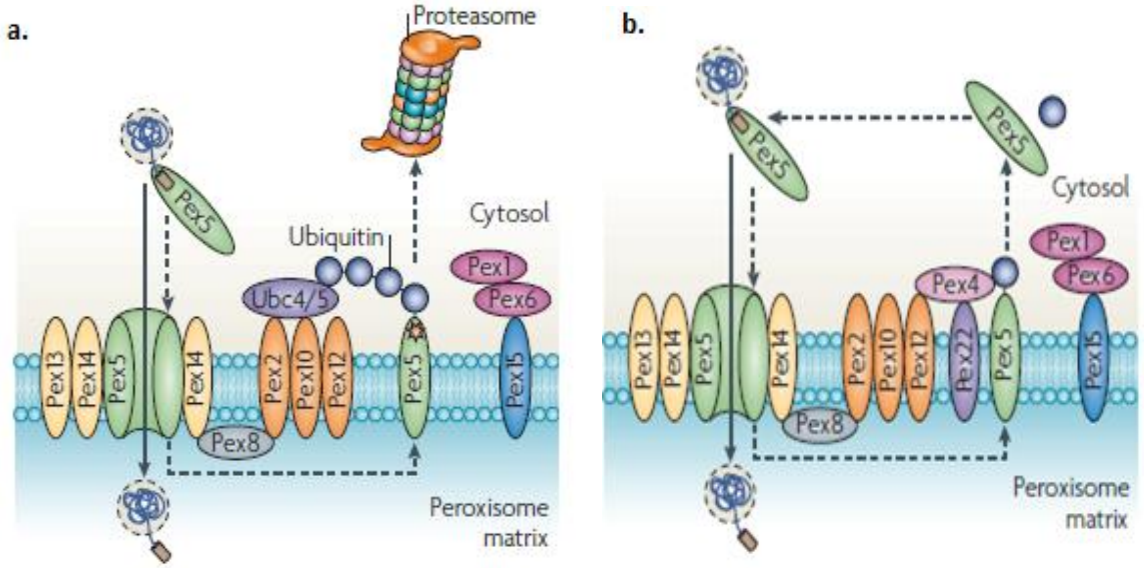


Figure 3. a. Quality control pathway where the receptor Pex5 is polyubiquitylated and eventually degraded by the proteasome **b.** Cycling pathway in which Pex5 is monoubiquitylated, released in the cytosol and available for the import cycle (Smith and Aitchison, 2013).

There are two pathways of peroxisome biogenesis that have been revealed using live-cell imaging with fluorescent reporters. Peroxisomes can be originated *de novo* by peroxisomal vesicles budding from the endoplasmic reticulum (ER) (Hoepfner et al., 2005) and then fuse with each other to create a mature peroxisome (van der Zand et al., 2012) or already existing peroxisomes can produce new peroxisomes through fission using new proteins and lipids that are provided from the ER as vesicles. The main difference between two pathways is that the *de novo* pathway has slower kinetics but produces peroxisomes that have all new material, while fission pathway is faster but necessitates the presence of already existing peroxisome (Motley and Hettema, 2007).

Current model that describes peroxisomal fission has four main steps: growth, elongation, constriction and fission (Smith and Aitchison, 2013). Firstly, mature peroxisome receives proteins and membrane from the ER through vesicular transport which allows peroxisome to grow. Then Pex11 protein activates and mediates tubulation of peroxisome

(Opaliński et al., 2010). The elongated membrane is enriched in membrane-anchored DRP (dynamin-related protein)-interacting proteins. Subsequently, the membrane is constricted and DRPs are recruited by DRP-interacting proteins (Motley, Ward and Hettema, 2008), which promote membrane scission to form new peroxisome (Schrader, Bonekamp and Islinger, 2012). It is important to note that DRPs and DRP-interacting proteins such as mitochondrial fission 1 (FIS1) also have established roles in mitochondrial fission (Losón et al., 2013).

In *de novo* peroxisome biogenesis, peroxisomes are formed by a second vesicular transport mechanism. Recently discovered model describes the production of two classes of pre-peroxisomal vesicles containing different sets of peroxisomal proteins which fuse together through pair-wise heterotypic fusion after they are released from the ER (van der Zand et al., 2012). Two vesicles separate RING finger and docking components of the import complex which prevents the import of matrix proteins until fusion and assembly of a functional import complex, *i.e.* forming mature peroxisome is finished (Smith and Aitchison, 2013).

1.4. Evolutionary origin of peroxisomes

All peroxisomes have highly conserved set of proteins that take part in protein import and organelle division which indicates single evolutionary origin. Nevertheless, they are identified in all major groups of eukaryotes, which mean they appeared before LECA. Since their discovery there were many assumptions about their evolutionary origin. First micrographs pointed a close relationship between the endoplasmic reticulum and the peroxisome (Novikoff and Shin, 1964), which initiated an idea about peroxisome being formed from the endoplasmic reticulum but that idea was fastly discarded in the scientific community. Instead, endosymbiotic origin was proposed because peroxisomes divide by fission from the pre-existing ones and import their proteins post-translationally, which is both a trait of organelles that have bacterial origin such as mitochondria and chloroplast (Lazarow and Fujiki, 1985).

de Duve suggested a phagocytic acquisition of a prokaryote that could detoxify oxygen through catalatic or peroxidatic mechanism and such bacterial originated organelle helped primitive anaerobic eukaryotes confront the rising concentration of oxygen in the atmosphere of the early anaerobic earth (de Duve, 1969). The absence of DNA in peroxisome was explained that it has entered the cell earlier than mitochondria and chloroplast, which only have vestigial amount of DNA compared to a bacterial genome (15×10^3 bp for human mitochondrial DNA/4.7×10^6 bp for *E.coli* DNA).

Theory about endosymbiotic origin of peroxisome has been widely accepted across scientific community for years, but in the last 20 years there were a lot of important discoveries which indicated different origin so the endosymbiotic theory about the peroxisomal origin was completly dismissed after those discoveries. Several experimental results pointed to a strong relationship between the endoplasmic reticulum and peroxisome. Peroxisome-less mutants in yeast can form new peroxisomes from endoplasmic reticulum when wild-type gene is introduced (Erdmann and Kunau, 1992). Using immuno-electron microscopy and electron tomography in mouse dendritic cells, it was shown that the peroxisomal membrane was derived from the ER (Tabak et al., 2003). As well, molecular machinery for the import of peroxisomal matrix proteins which is made of peroxins that are highy conserved among different eukaryotic taxa, shows homology to ER-associated protein degradation (ERAD) pathway (Gabaldón et al., 2006). ERAD machinery removes misfolded proteins from the lumen of ER and it is essential for the quality control of protein folding in ER. The similarity between the two machineries lies in the fact they make use of the same basic mechanistic principle: the tagging of a substrate by monoubiquitylation or polyubiquitylation, its subsequent recognition and ATP-dependent removal from a membrane by ATPases associated with diverse cellular activities (AAA) family of proteins (Schliebs, Girzalsky and Erdmann, 2010). Furthermore, phylogenetic analyses showed a common origin between peroxisomal proteins involved in protein import and the ones involved in the ERAD (Endoplasmic Reticulum Associated Decay) pathway (Gabaldón et al., 2006; Schluter et al., 2006), which would also imply that peroxisomes originated from the endoplasmic reticulum.

Nevertheless, mitochondrial origin of peroxisome without the ER involnment was also considered as a possible idea because of a high percentage of peroxiosmal enzymes that have been rerouted from mitochondria during evolution (Gabaldón and Pittis, 2015) and same factors are involved in mitochondrial and peroxiosmal fission (Losón et al., 2013). If mitochondrial origin of peroxiosme is true, it does not make much sense to secondarily replace mitochondrial protein import for an ER protein import system that is equally complicated. It seems more likely that the peroxisome secondarily adopted the dynamin-related fission from mitochondria allowing peroxisome to proliferate independently of the ER (Gabaldón, Ginger and Michels, 2016). Interestingly, last year it was discovered that within human patient fibroblasts lacking peroxisome newly synthesized peroxisomes are hybrids of mitochondrial and ER-derived pre-peroxisomes (Sugiura et al., 2017), which denies mithochondrial origin of peroxisome without the involvement of ER. Despite all the effort that was put in the research there is still an ongoing debate on the evolutionary origin of peroxisome.

1.5. Hypothesis and aim of study

The research goal of this master thesis is to address the question of the origin of peroxisome. In order to address this question a phylogenomic approach will be used based on the previously predicted ancestral eukaryotic peroxisomal proteome (Gabaldón et al., 2006) but using an expanded genomic dataset and state-of-the-art phylogenomic's techniques.

Minimal ancestral eukaryotic peroxisomal proteome contains twelve proteins of which six are peroxins (Pex1, Pex2, Pex4, Pex5, Pex10, Pex14), which are involved in protein import; three of them are linked to peroxisomal fatty acid β-oxidation (Fox1p, Fox2p, Faa2p), two of them take part in lipid transport (Pxa1p, Pxa2p), while the last one is the catalase (Cta1p) the hallmark peroxisomal protein which catalyzes the degradation of hydrogen peroxide. For each protein a phylogenetic tree will be reconstructed then complementing sub cellular localization predictions and protein domain information will be added to perform a detailed phylogenetic analysis which will have a goal to track the origin and other evolutionary events that could have shaped the emergence of peroxisome.

2. MATERIALS & METHODS

2.1. Database and query sequences

Database was built from eukaryotic and prokaryotic sequences. For the eukaryotic part 34 proteomes from 34 species were selected to include representatives of all major taxonomic groups whose members have their genome sequenced and publicly available. Proteomes from all main eukaryotic subdivisions were selected: 15 from Unikonts (13 Opisthokonta and 2 Amoebozoa), 8 from Chromalveolates, 6 from Plantae and 5 from Excavates (Supplementary 1.). All proteomes were retrieved from the NCBI Genome database (Release 224 February 15 2018) in the FASTA format.

For the prokaryotic part of the database the UniProt Reference Clusters 50 (UniRef50) for Bacteria and Archaea were retrieved from the UniProtKB (Release 2018_04). They were filtered to remove all the sequences that are environmental samples, uncultured samples, fragments or don't contain any taxonomic assignments. Nine archaeal proteomes from the Asgard group, which is considered to be the closest prokaryotic relative of eukaryotes (Eme et al., 2017), were also added to the database. Asgard proteomes were retrieved from the UniProt Proteomes database (Release 2018_02). In the end, collected FASTA files were used to make a BLAST database by using commands from the BLAST 2.2.31+ package. Final composition of the database is shown in the Table 1.

Table 1. Database composition according to the three domains of life.

Bacteria	Archaea	Eukaryota	Total
16,061,457	629,359	713,860	17,404,676

For each protein of the ancestral eukaryotic peroxisomal proteome orthologous protein sequences from human (*Homo sapiens)*, thale cress (*Arabidopsis thaliana*) and yeast (*Saccharomyces cerevisiae* strain ATCC 204508 / S288c) were selected as the query sequences for the BLAST search. The reasons why those three organisms were chosen is that they are all well studied, their proteomes are widely annotated and they are mutually phylogenetically remote which is convenient for detecting all homologs of the peroxisomal proteins in a broad range of species. All orthologs were retrieved from the UniProt database (Release 2018_02). If there were several orthologs from one species then the peroxisomal one was chosen. List of selected query sequences is available in Supplementary 2.

2.2. BLAST (Basic Local Alignment Search Tool) search

BLAST is an algorithm which compares a query sequence to the database to identify which database sequences resemble the query one above a certain threshold. It is a widely used bioinformatics tool to conclude functional and evolutionary relationships between sequences and to help identify members of gene families. The algorithm uses a heuristic approach which is based on locating short matches between sequences and after which local alignment is conducted. This approach makes the algorithm much faster than a full alignment procedure but less accurate. BLAST is a very practical solution because of its speed and relatively good accuracy comparing to full alignment procedures which are too slow for searching large databases (Mount, 2006).

Previously prepared protein database and protein query sequences were used as tool inputs for protein-protein BLAST (blastp) search. For each protein from the ancestral eukaryotic peroxisomal proteome three BLAST (version 2.2.31+) searches were performed with three different orthologs mentioned earlier as query sequences. Default parameters were used except for the E-value which was set on 0.001 and query coverage which was adjusted to 50%. Results from all three searches were combined and resulting protein sequences were used for building a profile HMM.

2.3. Profile HMM (Hidden Markov Model) search

Profile is defined as a consensus primary structure model consisting of position-specific residue scores and insertion or deletion penalties (Eddy, 1996). Instead of pairwise methods that use position-independent scoring, profile applies a position-specific scoring system to evaluate the degree of conservation at various positions which makes it a much more sensitive and specific method for database searching (Eddy, 1998).

HMM (Hidden Markov Model) is a finite model that describes a probability distribution over an infinite number of possible sequences (Eddy, 1996). Profile HMM is a linear state machine that has a series of nodes where each node roughly corresponds to a position in multiple alignment from which it was built, if we ignore the gaps then the correspondence is exact. While standard profile methods apply heuristic methods, profile HMMs have a formal probabilistic basis and apply consistant theory behind gap and insertion scores. A statistical method is used to estimate the true frequency from an observed frequency of a residue at a certain position in the alignment while standard profiles just apply the observed frequency to give a score for a certain position which makes a HMM profile a more advantageous method

(Eddy, 2018). Because of all mentioned above HMM profile search was used to detect evolutionary more distant homologs that cannot be identify by using pairwise methods such as BLAST.

Before building the HMM profiles, for each protein from the ancestral eukaryotic peroxisomal proteome multiple alignment was built from the BLAST result sequences. Multiple alignments were built using the program MAFFT v7.271 (Katoh et al., 2008) with the -auto parameter. HMM profiles were built from the alignments using the program HMMER (version 3.1b2) (Eddy, 2011). For the database search sequence e-value and domain e-value was 0.001 which is usually used as a significant value for distant homologs.

2.4. Sequence clustering

Sequence clustering based on sequence similarity was used when the total amount of hits in the HMM profile search was above 5000 or if a certain taxonomic group was overly represented (more than 90%). For that purpose Markov Cluster (MCL) algorithm was used which is a general purpose cluster algorithm for both weighted and unweighted networks (Enright, Van Dongen and Ouzounis, 2002).

Networks are structures in which nodes are connected by weighted links that are called edges. One node can be connected to multiple nodes while an edge connects two nodes. The nodes describe members from a defined class of objects such as genes or proteins and edges represent a weight that is a degree of similarity or dissimilarity between the nodes (objects) it connects. In this case the nodes are protein sequences while the edges are E-values that were obtained after the blastall search of query sequences against themselves.

Protocol 1. called "*Clustering protein sequence similarity networks*" from the book "*Bacterial Molecular Networks*" (van Dongen and Abreu-Goodger, 2011) was used for clustering and it will be briefly described. Firstly, blastall search was performed for each set of protein sequences that needed to be clustered. After that the following was done for each set: BLAST results were converted to ABC format, network file and label file were created then clustering was run several times with different values of inflation parameter (1.4, 2, 3, 4, 5, 6). Inflation parameter is the main parameter that affects the cluster granularity. Smaller values of inflation parameter give coarse grained clustering while higher values result in fine-grained clustering. The right choice of inflation parameter depends on the characteristics of the data so clustering should be run multiple times with different values of inflation parameter and then quality and coherency of the results should be tested.

Because of the above mentioned clustering outputs were compared and analyzed based on distance, efficiency and numerical criterion which describes granularity and capturing many edges in the input graph. When the best clustering output was chosen, one protein sequence was randomly selected from each cluster and taken for further analysis.

2.5. Reconstruction of phylogenetic trees

Phylogenetic trees were reconstructed from the set of homologous protein sequences that were prepared for each protein from the ancestral eukaryotic peroxisomal proteome. Phylogenetic pipeline from the PhylomeDB database was used for that purpose, which will be briefly described (Huerta-Cepas et al., 2010).

Firstly, protein sequences were aligned by using three different programs for multiple alignment: MAFFT v6.712b (Katoh et al., 2008), Muscle v3.7 (Edgar, 2004) and KAlign (Lassmann, Frings and Sonnhammer, 2008). Alignments were done in forward and reverse direction. From those six alignments (three alignments x 2 directions) consensus alignment was created by using M-Coffee (Wallace et al., 2006). Afterwards, consensus alignment was trimmed using the program TrimAl v1.2 (Capella-Gutierrez, Silla-Martinez and Gabaldón, 2009) to remove low consistency columns. Generated processed alignment was used to reconstruct phylogenetic trees using Neighbor Joining (NJ) and Maximum Likelihood (ML) methods. Following evolutionary models WAG, Blosum62, Dayhoff, JTT, LG, MtREV and VT were tested by evaluating the likelihood of the topology obtained by NJ allowing branch length optimization and the best one according to the AIC criterion was used for full ML approach. If the number of protein sequences was too high for using this pipeline then MAFFT v7.271 with –auto parameter and Fasttree 2.1.8 with –wag parameter (Price, Dehal and Arkin, 2010) was used instead.

2.6. Subcellular localization prediction

Subcellular localization prediction predicts where certain protein resides in the cell. Knowing the subcellular localization of the protein of interest is quite useful for experimental design, proteome analysis, genome annotation and identifying potential molecular targets for drugs (Meinken and Min, 2012). Here it will be used as a part of phylogenetic analysis to predict subcellular localization of proteins that are in the phylogenetic tree and to see how proteins from different subcellular compartments are related to peroxisomal proteins.

There are numerous computational tools for predicting the subcellular localization of eukaryotic proteins and most of them are based on one out of the two main concepts. First concept is based on the annotation of homologous sequences from knowledge databases. That means that for protein sequence of interest homologous sequences are found in the database that contains sequences with annotated subcellular localization. Second concept does not rely on any knowledge database yet it solely relies on the sequence itself. Sequence has regions that are important for its localization such as signal peptides, short motifs and sorting signals. Tools that are based on the second idea were chosen because they are more reliable for the purpose of this research. There are two reasons for that: first one is that for some proteins no annotated homologues sequences will exist so there will be no prediction of their localization and the second one is that there are a lot of sequence isoforms that share a high level of sequence similarity but have different subcellular localization and by using a method based on sequence similarity search they would end up having the same localization.

Requirements for predictors based on sequence similarity were that they can predict peroxisome as one of the subcellular localizations, can take all eukaryotic proteins as input not just certain taxonomic groups such as plants, animals, fungi etc. and to be high-throughput since there are a lot of proteins for analysis. Finally, chosen tools was DeepLoc (Almagro Armenteros et al., 2017). DeepLoc as a prediction model uses a recurrent neural network that processes the entire protein sequence and an attention mechanism for identifying protein regions important for the subcellular localization.

2.7. Domain annotation

Domain annotation for all proteins in each phylogenetic tree was done using HMM (Hidden Markov Model) profiles from the Pfam database (release 31.0). There were 16,712 HMM profiles where each profile describes one protein family (Punta et al., 2011). HMM profile search was done on the whole database with the e-value cut-off score 0,001. In the end each protein from the database was annotated and that was used further in the analysis.

2.8. Tree visualization

For the phylogenetic tree visualization iTOL 4.2.3 software (Letunic and Bork, 2016) was used. Interactive Tree of Life (iTOL) is web-based tool for display, manipulation and annotation of phylogenetic trees. Every phylogenetic tree was rooted using the outgroup method or midpoint rooting method, annotated according to taxonomic groups (Bacteria, Archaea, Asgard, Unikonts, Plantae, Chromalveolates, Excavates), subcellular localization and protein domains.

3. RESULTS

3.1. BLAST results

Results of BLAST search are shown in Table 1. and for each protein number of hits, their distribution across the three main domains of life and the highest e-value score is given. The number of hits varies from 17 to 670. All proteins involved in lipid metabolism have high number of hits varying from 399 as the lowest to 670 as the highest, while peroxins have quite lower number of hits varying from 17 as the lowest to 107 with Pex4 protein as an exception with 602 hits. Catalase is on the higher side of the spectrum with 334 hits. Also, the difference can be noticed in the distribution of hits across the main domains of life. Share of eukaryotic hits in peroxins ranges from 78.5% to 100% where Pex2, Pex10 and Pex14 have only eukaryotic hits. Pex1, Pex4 and Pex5 have 5.7%, 1.2%,14% of bacterial hits respectively as archaeal hits are the least represented. On the other side, proteins involved in lipid metabolism have the biggest share of bacterial hits which vary from 59.4% to 88.3% while eukaryotic hits range from 18.6% to 39.8% and archaeal hits are barely present in range from 0% to 1.9%. The highest e-value score is shown just to see the change of e-value after the query coverage filter. The highest e-value ranges from $7x10^{-3}$ to $3x10^{-42}$ which indicates intermediate sequence homology

Table 1. Results of BLAST search with e-value threshold 0.01 and query coverage 50%. For each protein distribution of hits across three main domains of life (Bacteria, Eukaryotes and Archaea) is given with the highest e-value score.

Protein	Number of hits	Eukaryotes	Bacteria	Archaea	The highest e-value
Cta1	334	62 (18.6%)	266 (79.6%)	6 (1.8%)	$7x10^{-3}$
Pxa1	448	118 (26.3%)	330 (73.7%)	0 (0%)	$9x10^{-14}$
Pxa2	401	117 (29.2%)	284 (70.8%)	0 (0%)	$7x10^{-10}$
Faa2	670	217 (32.2%)	444 (66.3%)	9 (1.5%)	$3x10^{-42}$
Fox1	399	159 (39.8%)	237 (59.4%)	3 (0.08%)	$3x10^{-3}$
Fox2	419	41 (9.8%)	370 (88.3%)	8 (1.9%)	$2x10^{-5}$
Pex1	52	44 (84.6%)	3 (5.7%)	5 (9,6%)	$1x10^{-7}$
Pex2	29	29 (100%)	0 (0%)	0 (0%)	$2x10^{-8}$
Pex4	602	592 (98.3%)	7 (1.2%)	3 (0.05%)	$8x10^{-9}$
Pex5	107	84 (78.5%)	15 (14.0%)	8 (7.5%)	$6x10^{-3}$
Pex10	37	37 (100%)	0 (0%)	0 (0%)	$5x10^{-3}$
Pex14	17	17 (100%)	0 (0%)	0 (0%)	$1x10^{-3}$

3.2. HMM profile search

Results of HMM profile search are presented in Table 2. which shows number of hits and its distribution across Eukaryota, Bacteria, Archaea and Asgard for each protein. Proteins involved in lipid metabolism have number of hits below e-value 0.001 in range from 3,127 to 85,387 which is too high to use them all for reconstructing a reliable phylogenetic tree. Since the share of bacterial hits s is extremely high, varies from 91.8% to 96.2%, they were clustered according to sequence similarity. Number of eukaryotic hits was very similar to number of archaeal hits so archaeal hits were also clustered according to sequence similarity because main point of interest is evolution of eukaryotic proteins.

Peroxins have lower number of hits and most of their hits are eukaryotic with Pex1 and Pex5 making an exception with 74.2% and 92.4% respectively and they also have the highest number of hits among all peroxins, so their bacterial and archaeal hits were also clustered according to sequence similarity for the same reason mentioned earlier.

Catalase has 885 hits and 775 of them are bacterial, 67 eukaryotic, 9 archaeal and 4 from Asgard group. Bacterial hits are clustered because there are too many of them to build a confident phylogenetic tree. Number of hits from Asgard group is the lowest in all proteins while Pex2 and Pex14 do not have any hits from Asgard group.

Table 2. Results of HMM profile search with the distribution of hits across domains Eukaryota, Bacteria, Archaea and Asgard.

Protein	Number of hits	Eukaryota	Bacteria	Archaea	Asgard
Cta1	855	67 (7.8%)	775 (90.6%)	9 (1.1%)	4 (0.5%)
Pxa1	85,387	2,454 (2.9%)	79,607 (93.2%)	2,937 (3.4%)	389 (0.5%)
Pxa2	83,950	2,393 (2.8%)	78,332 (93.3%)	2,841 (3.4%)	384 (0.5%)
Faa2	41,807	1,019 (2.4%)	40,203 (96.2%)	505 (1.2%)	80 (0.2%)
Fox1	9,132	501 (5.5%)	8,379 (91.8%)	192 (2.1%)	60 (0.6%)
Fox2	3,127	73 (2.3%)	2,936 (93.9%)	99 (3.2%)	19 (0.6%)
Pex1	8,896	1,740 (19.6%)	6,604 (74.2%)	485 (5.5%)	67 (0.7%)
Pex2	355	345 (97.2%)	6 (1.7%)	4 (1.1%)	0 (0.0%)
Pex4	1,318	1,207 (91.5%)	78 (6.0%)	17 (1.3%)	16 (1.2%)
Pex5	66,653	3,118 (4.7%)	61,616 (92.4%)	1,802 (2.7%)	117 (0.2%)
Pex10	3,006	2,876 (95.7%)	115 (3.8%)	13 (0.4%)	2 (0.1%)
Pex14	44	36 (81.8%)	7 (16.0%)	1 (2.2%)	0 (0.0%)

3.3. Clustering results

Clustering based on sequence similarity was conducted separately for bacterial and archaeal protein sequences. Clustering of bacterial sequences was done for Cta1, Pxa1, Pxa2, Faa2, Fox1, Fox2, Pex1 and Pex5, while the clustering of archaeal sequences was done for Pxa1, Pxa2, Faa2, Fox1, Fox2, Pex1 and Pex5. The right choice of inflation value as a main parameter in Markov clustering algorithm (MCL) depends on the characteristics of the data so the quality and coherency of each clustering was tested. For set of inflation values 1.4, 2, 3, 4, 5 and 6 number of clusters, cluster distance between each two consecutive clusterings and performance measures-efficiency, mass fraction, area fraction and jury pruning synopsis are presented and taken into account when selecting the best clustering. Cluster distance is calculated as the number of nodes required to change location in one clustering in order to obtain the other is divided with the total number of nodes, i.e. total number of sequences. Lower fraction of nodes that require relocation between two clusters mark higher stability which means that clusters represent sets of highly conserved sequences that do not easily split further. Efficiency factor is a measurement that takes value in range 0-1 and achieves 1 only when natural clustering exists which means all nodes within one cluster are connected with edges while there is no edge connection between nodes from different clusters. The mass fraction is the joint edge weight of all captured edges divided by the joint weight of all edges. Captured edge is an edge between two nodes that are in the same cluster. The area fraction is roughly the sum of squares of all cluster sizes divided by the square of the number of nodes in the graph. Lower or higher area fraction indicates more granulated or coarser clustering. Jury pruning synopsis indicates how well the pruning went, i.e. if the scheme parameter is high enough. Scheme parameter lies in range 1-7 and the highest value was used for all clustering.

3.3.1. Bacteria

In Table 3. number of clusters for set of inflation values 1.4, 2, 3, 4, 5 and 6 are shown for proteins Cta1, Pxa1, Pxa2, Faa2, Fox1, Fox2, Pex1 and Pex5. As the inflation value gets higher number of clusters increases which is expected because inflation value affects cluster granularity in a way that higher values will result in more fine-grained clustering as lower values give coarse grained clustering.

In Table 4. cluster distance between two consecutive clustering are shown. Proteins Cta1, Fox1, Fox2 and Pex1 show very low cluster distance between clustering with inflation values 5 and 6. Pxa1 and Pxa2 have the lowest cluster distance between clusters made with inflation values 3 and 4. For protein Faa2 cluster distances vary from 13,00% to 18,46%

while for Pex5 cluster distance falls as the inflation value increases but number of clusters at inflation value 5 starts to outnumber the number of eukaryotic hits and their overall number is still too high to reconstruct a phylogenetic tree. Proteins Pxa1, Pxa2, Pex5 and Faa2 have high cluster distance between their clusterings which is expected because they have very high number of bacterial hits that probably share lower similarity.

In Table 5. efficiency, mass fraction, area fraction and jury pruning synopsis for each clustering are presented. Proteins Cta1, Fox2 and Pex1 have the highest efficiency values that are above 0.7 for certain inflation values which indicates that clustering is granular and captures many edges in the input graph at same time. Pex5 and Fox1 have efficiency values around 0.5 for higher inflation values which is a decent performance considering the high number of sequences that are clustered. Pxa1, Pxa2 and Faa2 have the lowest efficiency values around 0.3 for higher inflation values which is expected because they have the highest number of bacterial sequences that probably do not represent set of highly conserved sequences.

All proteins have high results for mass fraction. Knowing that mass fraction decreases with inflation value. Proteins Cta1, Pxa1, Pxa2, Fox2, Pex1 have mass fraction between 0.73 and 0.82 when the inflation value is set on 5 which is a high result considering such a high inflation value. Faa2 and Fox1 have slightly lower results, 0.67 and 0.60 respectively while Pex5 has the lowest 0.5 when the inflation value is 5.

Area fraction indicates the granularity of the clustering. Pxa1, Pxa2 and Pex5 have the lowest results that lay between 0.06 and 0.02 which indicates very fine granularity. For Pex1 and Fox2 score is around 0.1 when the inflation value is set to 5 while Faa2 and Fox1 are around 0.15. Cta1 has the highest area fraction 0.24 when the inflation value is 5 which mean that Cta1 has the coarsest clustering across all inspected proteins.

For proteins Cta1, Fox1, Fox2 and Pex1 jury pruning synopsis is above 80 which means that scheme parameter is high enough while proteins Pxa1, Pxa2, Faa2 and Pex5 have values around 50 which is considered to be tolerable even though the highest possible value of scheme parameter is used.

For proteins Cta1, Fox1, Fox2, Faa2 and Pex1 clustering with inflation value 5 is chosen because they show the highest cluster stability based on calculated cluster distance , favorable mass fraction and high granularity. For proteins Pxa1, Pxa2 and Pex5 inflation value 3 is chosen because they have too high granularity at inflation value 5 to the extent that the number of bacterial clusters will outnumber number of eukaryotic hits considering that one sequence is randomly chosen from each cluster.

Table 3. Number of clusters of bacterial sequences for set of inflation values 1.4, 2, 3, 4, 5 and 6 for proteins Cta1, Pxa1, Pxa2, Faa2, Fox1, Fox2, Pex1 and Pex5.

Inflation value	1.4	2	3	4	5	6
CTA1	3	5	12	15	16	16
PXA1	35	77	188	542	1,375	2,610
PXA2	33	74	176	497	1,268	2,461
FAA2	8	38	79	101	235	525
FOX1	5	14	28	43	54	64
FOX2	7	18	30	43	53	65
PEX1	16	35	51	63	68	73
PEX5	371	1,115	1,822	2,325	2,935	3,570

Table 4. Cluster distance between consecutive clusterings of bacterial sequences sequences for pairs of inflation values 1.4/2, 2/3, 3/4, 4/5 and 5/6 for proteins Pxa1, Pxa2, Faa2, Fox1, Fox2, Pex1 and Pex5.

Clusters compared	1.4/2	2/3	3/4	4/5	5/6
CTA1	32,77%	7,22%	9,93%	2,83%	0,51%
PXA1	21,77%	9,76%	6,04%	8,26%	14,07%
PXA2	21,55%	11,11%	6,58%	7,84%	14,43%
FAA2	13,00%	17,39%	15,15%	16,29%	18,46%
FOX1	15,03%	11,14%	6,14%	2,11%	1,68%
FOX2	31,16%	5,27%	17,54%	18,46%	4,73%
PEX1	18,64%	6,05%	2,74%	1,71%	1,74%
PEX5	38,86%	25,90%	14,71%	10,72%	8,22%

Table 5. Performance measures efficiency, mass fraction, area fraction and jury pruning synopsis are presented for set of inflation values 1.4, 2, 3, 4, 5 and 6 for clustering of bacterial sequences of proteins Cta1, Pxa1, Pxa2, Faa2, Fox1, Fox2, Pex1 and Pex5.

Protein	Inflation value	Efficiency	Mass fraction	Area fraction	Jury pruning synopsis
Cta1	1.4	0.50571	0.97682	54.413	98.8 or marvelous
	2	0.70518	0.92343	31.753	99.0 or perfect
	3	0.74489	0.85731	28.062	99.0 or perfect
	4	0.74568	0.77354	24.556	99.0 or perfect
	5	0.74163	0.77038	24.599	99.0 or perfect
	6	0.74068	0.77071	24.634	99.0 or perfect
Pxa1	1.4	0.20540	0.91481	6.668	40.4 or bad
	2	0.29224	0.84495	4.281	45.0 or dodgy
	3	0.32715	0.80470	3.604	49.1 or so-so
	4	0.34127	0.78143	3.324	51.4 or mediocre
	5	0.34510	0.73113	2.960	52.2 or mediocre
	6	0.35711	0.65581	2.312	52.8 or mediocre
Pxa2	1.4	0.20798	0.91564	6.675	40.6 or bad
	2	0.29373	0.84699	4.333	45.2 or dodgy
	3	0.32118	0.80935	3.711	49.1 or so-so
	4	0.33818	0.78709	3.417	51.4 or mediocre
	5	0.34147	0.74125	3.062	52.2 or mediocre
	6	0.35698	0.66477	2.385	52.8 or mediocre
Faa2	1.4	0.08117	0.97207	42.386	46.4 or shoddy
	2	0.15744	0.91667	35.566	49.6 or so-so
	3	0.21380	0.82293	25.227	52.8 or mediocre
	4	0.23848	0.76524	21.730	53.8 or tolerable
	5	0.27504	0.67972	15.754	54.4 or tolerable
	6	0.28555	0.58167	11.301	54.6 or tolerable
Fox1	1.4	0.28656	0.96082	32.494	78.0 or groovy
	2	0.41819	0.90156	23.197	79.0 or not bad at all
	3	0.47541	0.83991	19.145	80.0 or favourable
	4	0.48862	0.79723	17.915	80.5 or favourable
	5	0.48696	0.78500	17.649	80.8 or favourable
	6	0.48848	0.77327	17.154	81.0 or cracking
Fox2	1.4	0.55122	0.83054	25.462	90.0 or cracking
	2	0.60567	0.80425	20.749	90.9 or cracking
	3	0.63039	0.77888	19.384	91.6 or cracking
	4	0.68755	0.67348	13.181	92.0 or scrumptious
	5	0.71446	0.60993	9.726	92.2 or scrumptious
	6	0.71992	0.60580	9.440	92.5 or scrumptious
Pex1	1.4	0.51572	0.93626	22.647	86.2 or cracking
	2	0.66356	0.89024	13.298	87.2 or cracking
	3	0.70260	0.84894	10.873	88.2 or cracking
	4	0.71522	0.83045	10.164	88.5 or cracking
	5	0.71955	0.82171	9.782	88.8 or cracking
	6	0.72429	0.81054	9.385	88.8 or cracking
Pex5	1.4	0.33422	0.75433	14.780	41.8 or bad
	2	0.45678	0.64270	5.201	46.0 or shoddy
	3	0.48603	0.58923	4.526	50.1 or mediocre
	4	0.52127	0.53757	2.449	51.8 or mediocre
	5	0.52807	0.50928	2.039	52.2 or mediocre
	6	0.53375	0.49053	1.832	52.8 or mediocre

3.3.2. Archaea

Table 6. presents number of clusters of archaeal sequences for set of inflation values 1.4, 2, 3, 4, 5 and 6. For all proteins it can be noticed that at inflation values 5 and 6 number of clusters stabilize and there is not a big difference anymore. This is also confirmed with cluster distances between consecutive clusterings that are shown in Table 7. For all proteins clusterings with the highest inflation values have cluster distance in range 0,41% to 6,33% which indicates high stability of the clustering. For inflation values 5 and 6 proteins Faa2, Fox1, Fox2 and Pex1 have mass fraction above 0.7, Pxa1 and Pxa2 are around 0.55 while Pex5 is around 0.5 which are all favorable results. Area fraction for inflation values 5 and 6 of all proteins lies in range from 0.18 to 0.82, which indicates very coarse granularity but that can be noticed from the small number of clusters that are found for such a high parameter of inflation values. All proteins have a jury pruning synopsis above 80, which means that scheme parameter settings are high enough. Comparing to clustering of bacterial sequences archaeal clustering shows better results in stability and jury pruning synopsis because number of archaeal sequences to cluster was much lower. Finally, for all proteins clustering with inflation value 5 was chosen because it shows almost the same stability, number of clusters and performance results as the highest inflation value that was tested.

Table 6. Number of clusters of archaeal sequences for set of inflation values 1.4, 2, 3, 4, 5 and 6 for proteins Pxa1, Pxa2, Faa2, Fox1, Fox2, Pex1 and Pex5.

Inflation value	1.4	2	3	4	5	6
Pxa1	1	5	9	12	18	23
Pxa2	1	5	9	12	18	23
Faa2	2	3	8	8	10	10
Fox1	1	2	2	2	2	3
Fox2	2	3	3	3	3	3
Pex1	5	7	11	11	12	12
Pex5	6	20	41	48	60	62

Table 7. Cluster distances between consecutive clusterings of archaeal sequences for pairs of inflation values 1.4/2, 2/3, 3/4, 4/5 and 5/6 for proteins Pxa1, Pxa2, Faa2, Fox1, Fox2, Pex1 and Pex5.

Clusters compared	1.4/2	2/3	3/4	4/5	5/6
Pxa1	45,39%	13,11%	11,47%	4,90%	4,80%
Pxa2	44,64%	12,86%	11,93%	4,45%	5,17%
Faa2	0,99%	34,85%	8,71%	6,93%	6,33%
Fox1	6,25%	2,08%	0,00%	0,00%	2,60%
Fox2	21,21%	2,02%	0,00%	0,00%	2,02%
Pex1	14,43%	6,59%	0,41%	1,64%	0,41%
Pex5	12,86%	6,51%	2,69%	3,08%	1,01%

Table 8. Performance measures efficiency, mass fraction, area fraction and jury pruning synopsis are presented for set of inflation values 1.4, 2, 3, 4, 5 and 6 for clustering of archaeal sequences of proteins Pxa1, Pxa2, Faa2, Fox1, Fox2, Pex1 and Pex5.

Protein	Inflation value	Efficiency	Mass fraction	Area fraction	Jury pruning synopsis
Pxa1	1.4	0.20530	1.00000	1.00000	83.1 or cracking
	2	0.48434	0.77399	0.37967	84.5 or cracking
	3	0.53964	0.69558	0.27127	85.5 or cracking
	4	0.58069	0.65080	0.20354	86.1 or cracking
	5	0.57499	0.62084	0.18801	86.4 or cracking
	6	0.56983	0.60494	0.18093	86.6 or cracking
Pxa2	1.4	0.21165	1.00000	1.00000	84.0 or cracking
	2	0.48833	0.77215	0.38436	85.1 or cracking
	3	0.54247	0.69381	0.27510	86.1 or cracking
	4	0.58467	0.64834	0.20533	86.8 or cracking
	5	0.57843	0.61953	0.19001	87.0 or cracking
	6	0.56772	0.59856	0.18304	87.2 or cracking
Faa2	1.4	0.55597	0.99076	0.96492	99.0 or perfect
	2	0.57077	0.98193	0.94563	99.0 or perfect
	3	0.71266	0.65907	0.46313	99.0 or perfect
	4	0.71595	0.66007	0.46483	99.0 or perfect
	5	0.71004	0.59350	0.39728	99.0 or perfect
	6	0.70873	0.57422	0.37258	99.0 or perfect
Fox1	1.4	0.63706	100.000	1.00000	100.0 or really good
	2	0.70696	0.94163	0.88220	99.6 or perfect
	3	0.71754	0.93593	0.86409	99.6 or perfect
	4	0.71754	0.93593	0.86409	99.6 or perfect
	5	0.71754	0.93593	0.86409	99.6 or perfect
	6	0.71621	0.92894	0.86164	99.6 or perfect
Fox2	1.4	0.64702	0.94246	0.81653	99.6 or perfect
	2	0.82588	0.83501	0.52216	99.6 or perfect
	3	0.82192	0.82717	0.51268	99.6 or perfect
	4	0.82192	0.82717	0.51268	99.6 or perfect
	5	0.82192	0.82717	0.51268	99.6 or perfect
	6	0.82110	0.82034	0.50361	99.6 or perfect
Pex1	1.4	0.51025	0.97451	0.89634	99.0 or perfect
	2	0.69721	0.91700	0.66545	99.0 or perfect
	3	0.74465	0.86009	0.58083	99.0 or perfect
	4	0.74631	0.85854	0.57776	99.0 or perfect
	5	0.75504	0.84231	0.55710	99.0 or perfect
	6	0.75635	0.84037	0.55414	99.0 or perfect
Pex5	1.4	0.29263	0.99358	0.98100	94.9 or ripping
	2	0.43070	0.91461	0.75311	94.9 or ripping
	3	0.47106	0.86379	0.68864	95.0 or fabulous
	4	0.47828	0.84667	0.67295	95.0 or fabulous
	5	0.48539	0.82714	0.65705	95.2 or fabulous
	6	0.48560	0.82430	0.65507	95.2 or fabulous

3.4. Reconstruction of phylogenetic trees

Firstly, for each protein from the ancestral eukaryotic peroxisomal proteome two phylogenetic trees were reconstructed. First phylogenetic tree was reconstructed from the protein sequences detected by the HMM profile (Table 2.) that was built from the sequences that were found in BLAST search (Table 1.). This set of phylogenetic trees was named eukaryotic trees because they also contain other eukaryotic protein families besides the peroxisomal one.

Second phylogenetic tree was reconstructed from the sequences that were detected by the HMM profile built from the orthologous eukaryotic protein sequences found in the MetaPhors database (Pryszcz, Huerta-Cepas and Gabaldón, 2010) and because of that this set of phylogenetic trees was named phylogenetic trees based on orthology search.

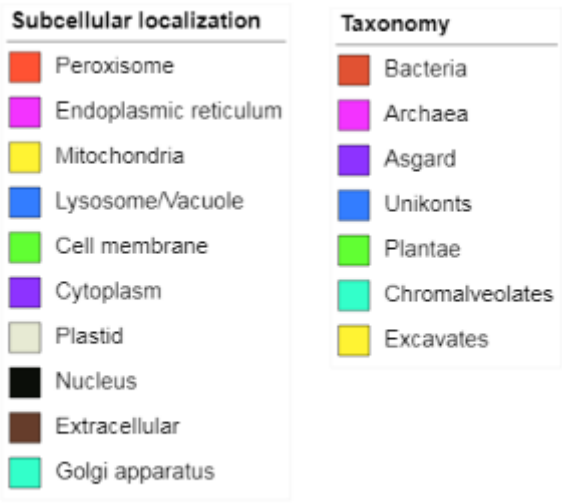
Third set of phylogenetic trees was reconstructed only for the proteins whose eukaryotic trees were excessively large. Because this set of trees was derived out of the peroxisomal subtrees of the eukaryotic trees, they were named peroxisomal trees.

All trees are circularly visualized and mostly contain four datasets that represent taxonomy, subcellular localization, bacterial phyla and protein domain composition. First circle contains unique code for each protein which is made of TaxID from the NCBI Taxonomy database of the organism the protein is from as a first value and the ordinal number of the protein in the proteome of the organism as the second value. Each code is colored according to the taxonomy of the organism to which each protein belongs. Prokaryotes are divided into Bacteria, Archaea an Asgard group while Eukaryotes are divided into Unikonts, Plantae, Chromalveolates and Excavates. Particular color used for each taxonomic group is shown in Legend 1.

Second circle presents subcellular localization of eukaryotic proteins. Possible locations are nucleus, cytoplasma, extracellular, cellular membrane, peroxisome, endoplasmic reticulum, mitochondrion, chloroplast, Golgi apparatus and lysosome/vacuole. List of colors used for each subcellular location are shown in Legend 1.

Third circle shows phyla to which bacterial protein sequences belong. This is shown because most of the trees had bacterial proteins in several places and since they were used as an outgroup to root the tree knowing the phyla made it sometimes more obvious which bacterial protein should be used as an outgroup. List of colors used for each bacterial phylum is shown in Legend 2. Final circle presents protein domain composition of each protein. All protein domains are shown as rectangular boxes of the same size but different color. Colors of the most important and common domains are mentioned in the description of each tree.

Legend 1. List of colors used for taxonomic information and subcellular localization prediction. First circle shows taxonomic information for each protein while second circle depicts predicted subcellular localization of eukaryotic proteins.



Legend 2. List of colors used for each bacterial phylum. Third circle in each phylogenetic tree shows phyla to which bacterial protein sequences belong.

3.4.1. Eukaryotic phylogenetic trees

Distribution of sequences used in the reconstruction of this set of phylogenetic trees across different domains of life is shown in Table 12. Archaea is divided into Asgard which represents Asgard group and Archaea that contains all other archaeal species. For proteins Cta1, Pxa1, Pxa2, Fox1, Fox2, Faa2, Pex1 and Pex5 clustering of bacterial and archaeal sequences was done (see 3.3. Clustering results) and one representative of each cluster was randomly selected and used for the reconstruction of the tree. For proteins Pex10 that mostly had eukaryotic hits first 500 results with the highest e-value from the HMM search were used because the total amount of detected protein sequences was too high to build a reliable phylogenetic tree and there was no sense to look for a peroxisomal subtree due to low number of bacterial hits since the Pex10 is considered to have a eukaryotic origin (Gabaldón et al, 2006). For proteins Cta1, Fox1, Fox2, Pex10 and Pex14 phylogenetic pipeline was used while for other proteins due to high number of protein sequences MAFFT with the –auto parameter were used to built the alignment and Fasttree with the –wag parameter to reconstruct the tree.

For proteins Pxa1, Pxa2, Faa2, Pex1 and Pex5 that have excessively large trees, it was decided to find a peroxisomal subtree that would be cut and rebuild using a more reliable pipeline. Since those trees contain several eukaryotic protein families, peroxisomal family was determined in a way that firstly the clade that contains query sequences from *Homo sapiens*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae* for whom it is known that are peroxisomal was found and then branches would be added until representatives of all four major eukaryotic groups were captured and prokaryotic proteins appeared so that was considered as LECA (Last Eukaryotic Common Ancestor) of that peroxisomal family.

Table 9. Distribution of protein sequences used for the reconstruction of eukaryotic trees across domains Eukaryota, Bacteria, Archaea and Asgard.

Protein	Number of hits	Eukaryota	Bacteria	Archaea	Asgard
Cta1	96	67	16	9	4
Pxa1	3,111	2,454	188	18	451
Pxa2	3,030	2,393	176	18	443
Faa2	1,344	1,019	235	10	80
Fox1	617	501	54	2	60
Fox2	148	73	53	3	19
Pex1	1,887	1,740	68	12	67
Pex2	355	345	6	4	0
Pex4	1,318	1,207	78	17	16
Pex5	3,688	3,118	371	60	139
Pex10	500	484	15	1	0
Pex14	44	36	7	1	0

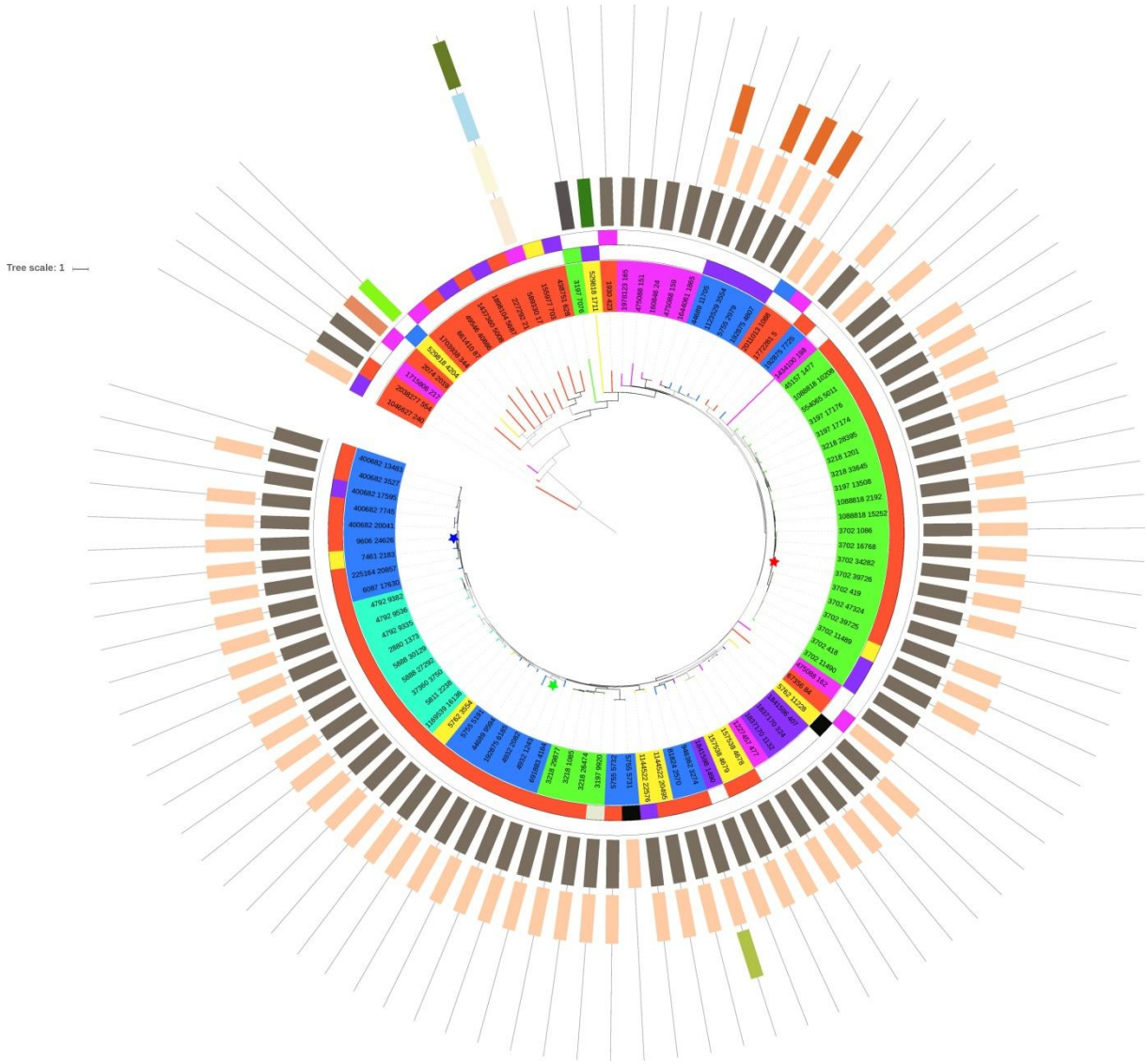


Figure 4. Phylogenetic tree of the protein catalase (Cta1) reconstructed from the protein sequences that were detected by HMM profile that was built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 12 different domains were detected and two most common ones are catalase domain shown in grey and catalase-related immune responsive domain shown in beige color that are found in eukaryotic, bacterial and archael protein sequences. Tree was rooted using an outgroup method.

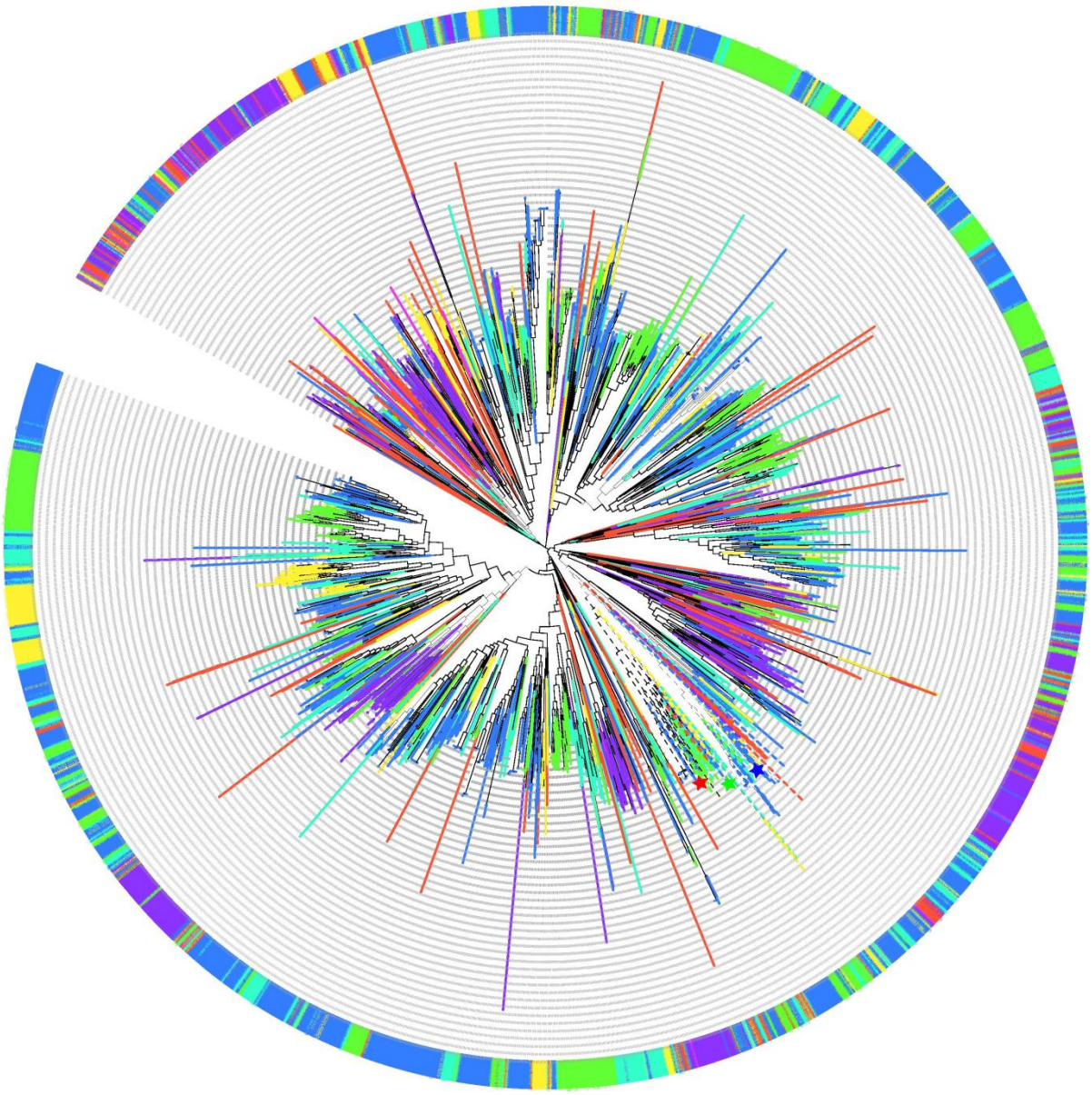


Figure 5. Phylogenetic tree of the peroxisomal long-chain fatty acid import protein 2 (Pxa1) reconstructed from the protein sequences that were detected by HMM profile that was built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Peroxisomal clade is labeled with dashed lines.

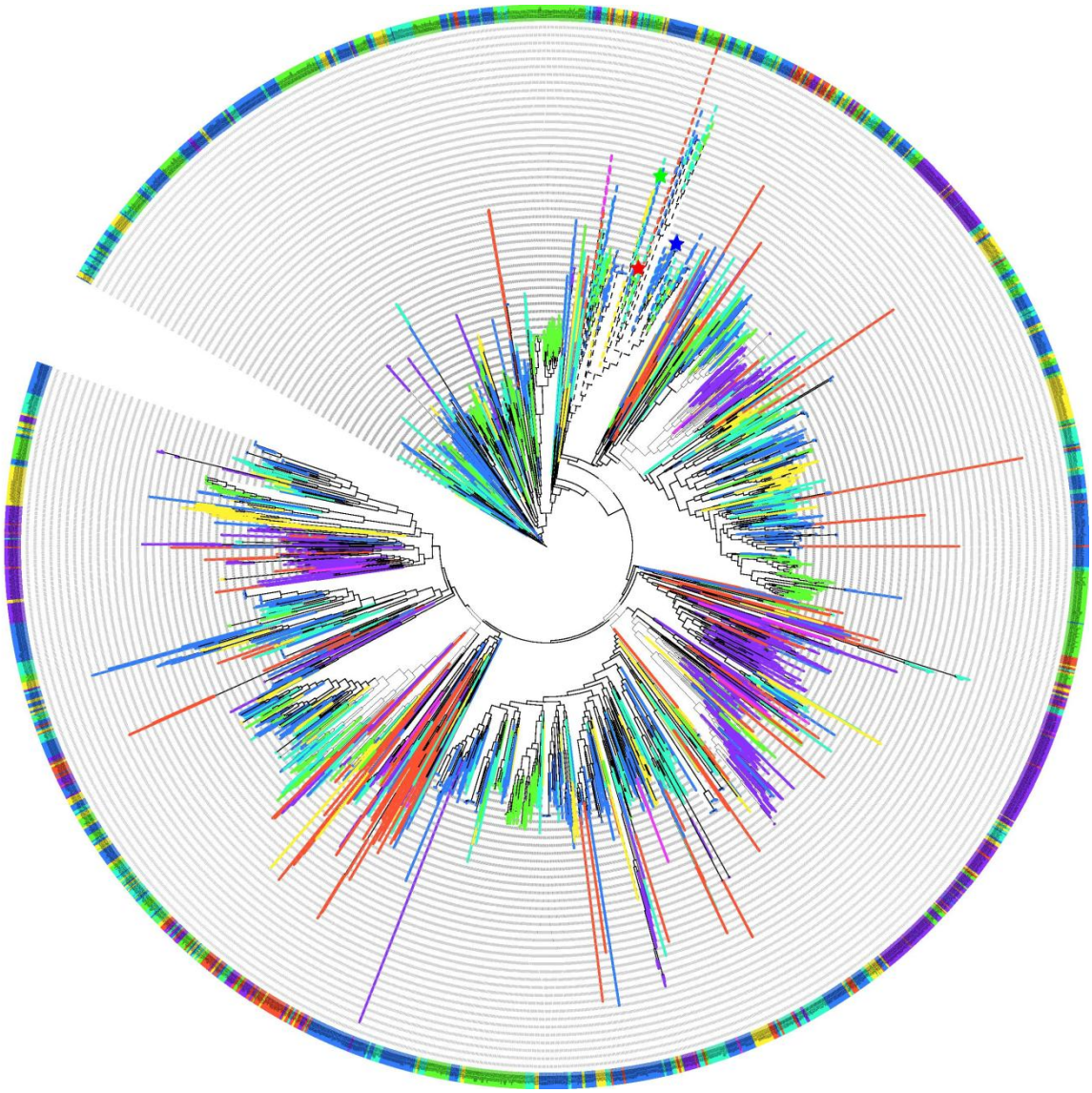


Figure 6. Phylogenetic tree of the Peroxisomal long-chain fatty acid import protein 1 (Pxa2) reconstructed from the protein sequences that were detected by HMM profile that was built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Peroxisomal clade is labeled with dashed lines.

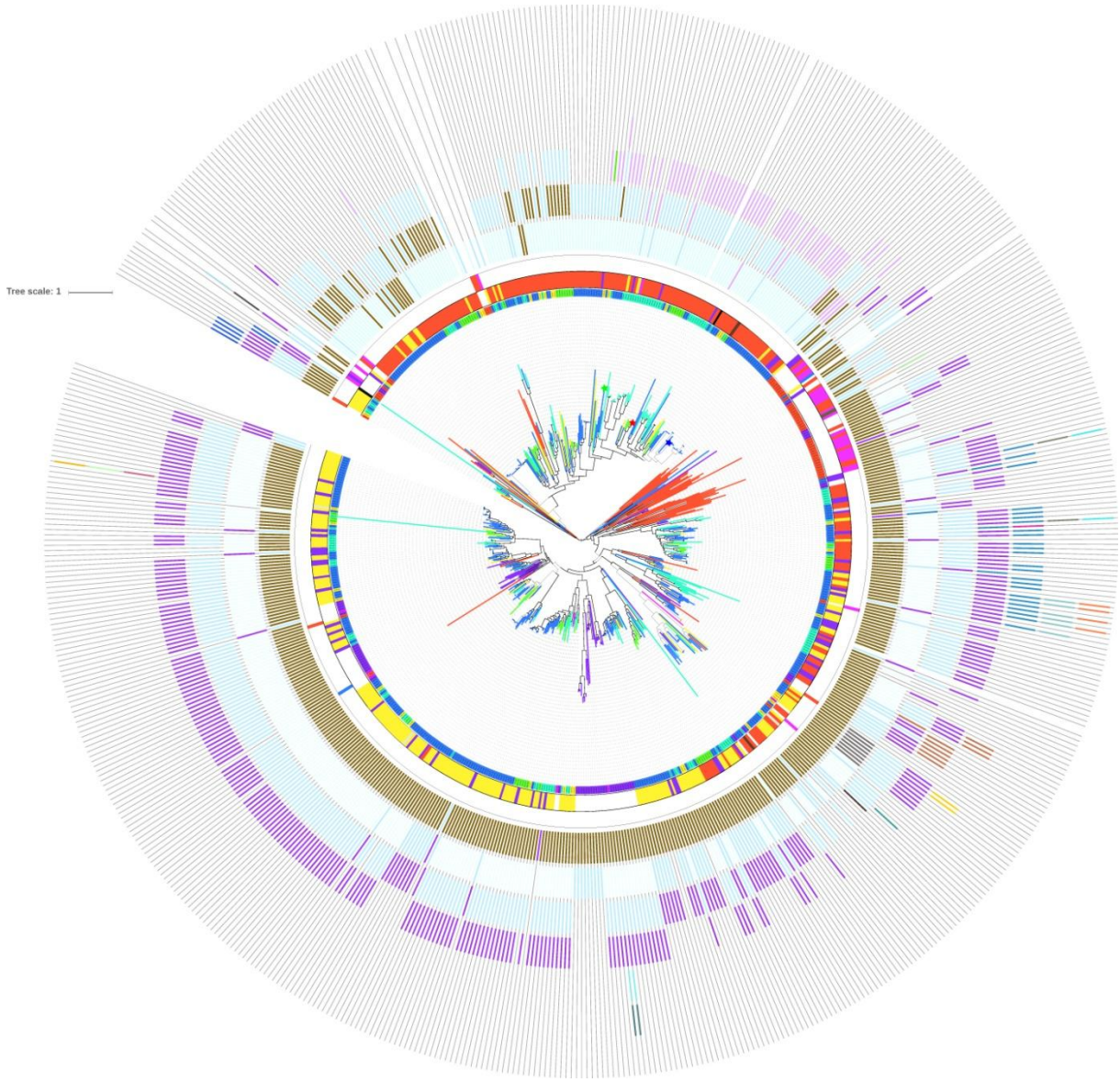


Figure 7. Phylogenetic tree of the Acyl-coenzyme A oxidase (Fox1) reconstructed from the protein sequences that were detected by HMM profile that was built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Tree was rooted using a midpoint method. In total 28 domains were detected and the most common are Acyl-CoA dehydrogenase middle domain, Acyl-CoA dehydrogenase C-terminal domain 1, Acyl-CoA dehydrogenase N-terminal domain, Acyl-CoA dehydrogenase C-terminal domain 2, Acyl-CoA oxidase domain and Acyl-CoA oxidase N-terminal domain and are shown in light blue, oak brown, purple, azure, light cyan and mauve respectively.

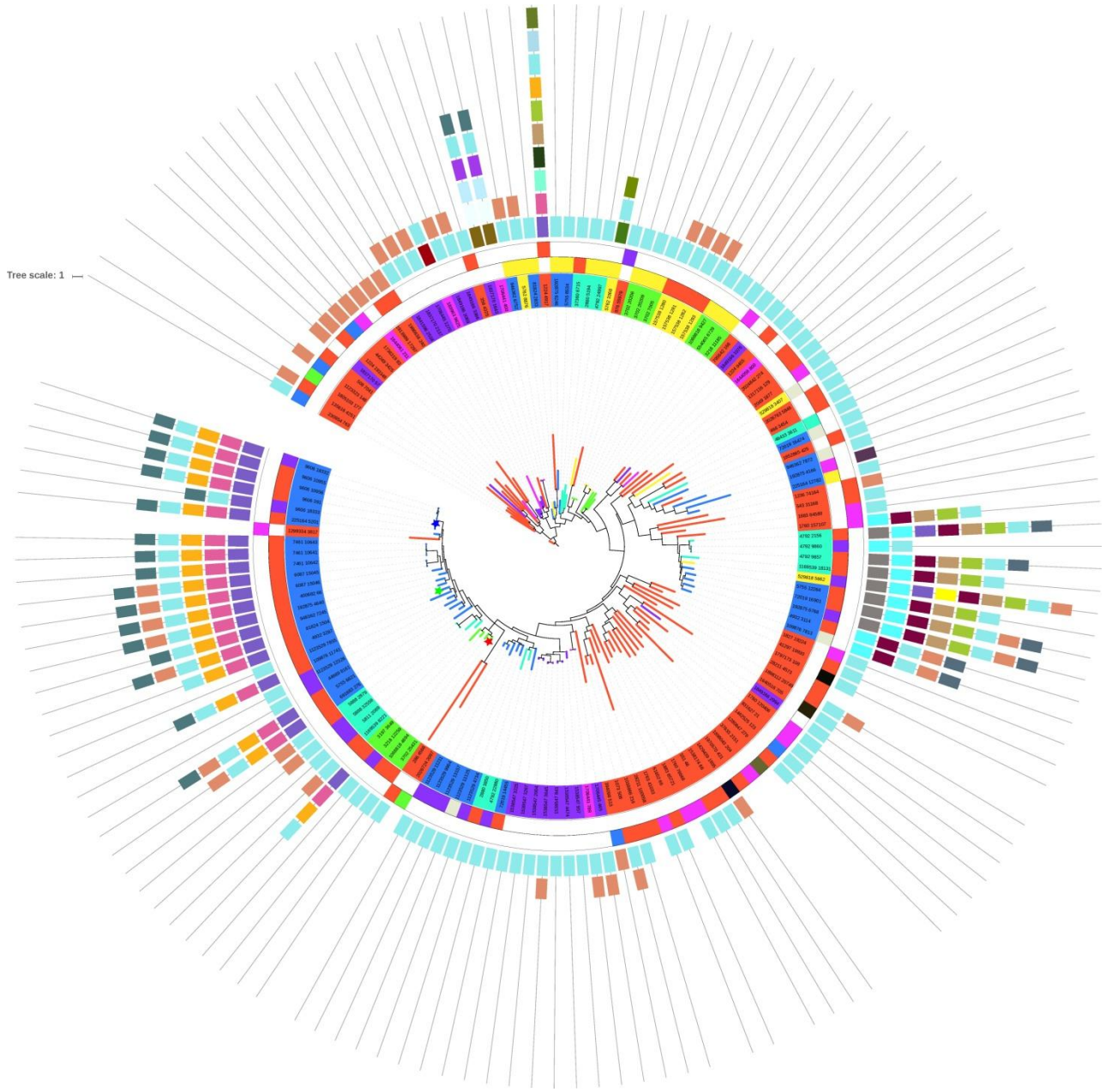


Figure 8. Phylogenetic tree of the Peroxisomal hydratase-dehydrogenase-epimerase (Fox2) reconstructed from the protein sequences that were detected by HMM profile that was built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Tree was rooted using an outgroup method. In total 25 different domains are detected and the most common are MaoC like domain, N-terminal half of MaoC dehydratase, short chain dehydrogenase, Enoyl-(Acyl carrier protein) reductase, KR domain, SCP-2 sterol transfer family and are shown in blue lagoon, dark salmon, purple sage bush, pink cupcake, beer, beetle green respectively. Eukaryotic sequences can roughly be divided into three clades – first one contains mostly mitochondrial protein sequences from all four major groups of eukaryotes, second one covers peroxisomal and cytoplasmic protein sequences from Excavates, Chromalveolates and Unikonts, third clade contains all three query sequences and peorxisomal sequences from Unikonts, Plantae and Chromalveolates. Interestingly, eukaryotic proteins from all three clades show different protein domain composition.

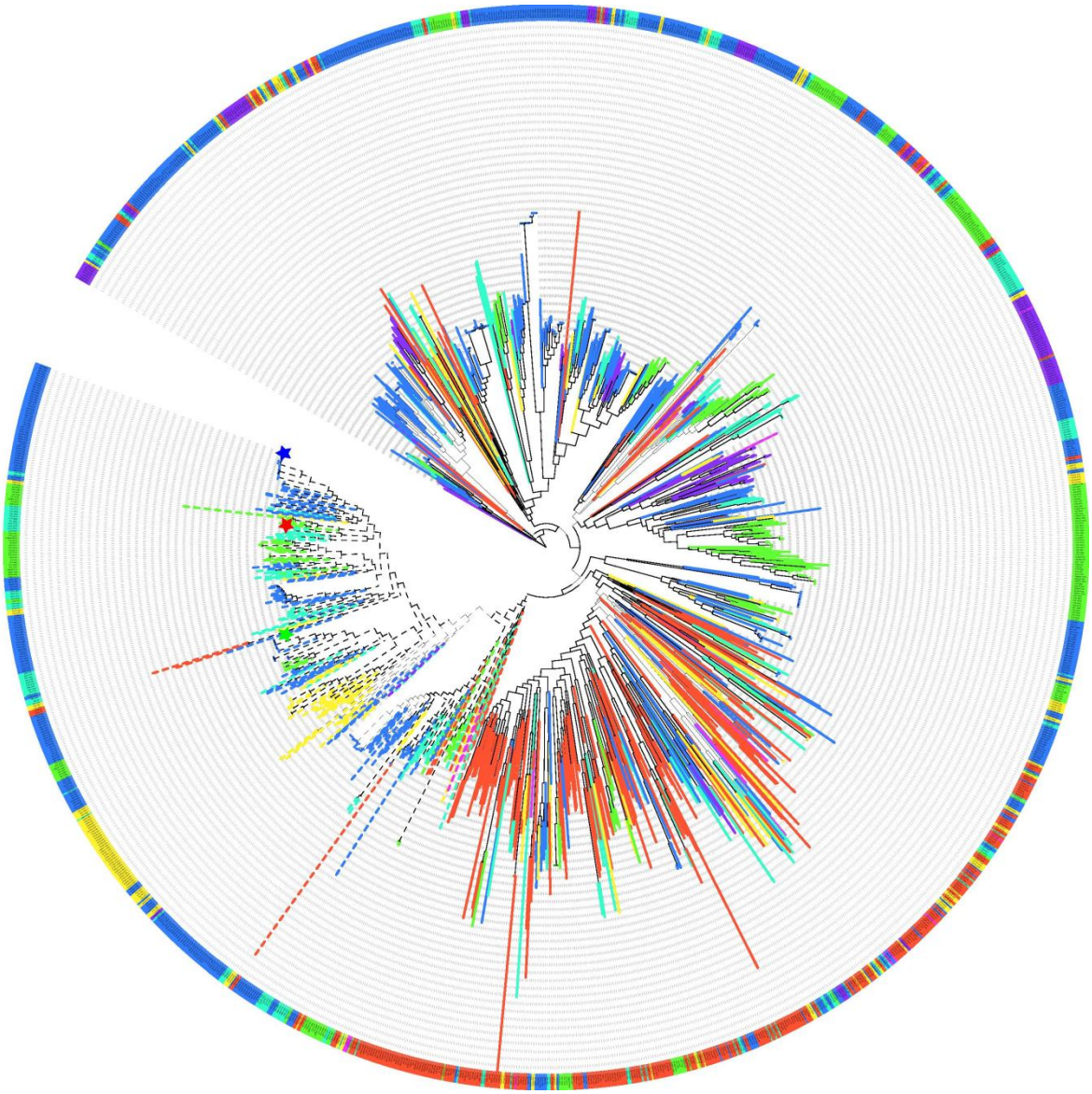


Figure 9. Phylogenetic tree of the Long-chain-fatty-acid-CoA ligase 2 (Faa2) reconstructed from the protein sequences that were detected by an HMM profile that was built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Selected peroxisomal clade is labeled with dashed lines.

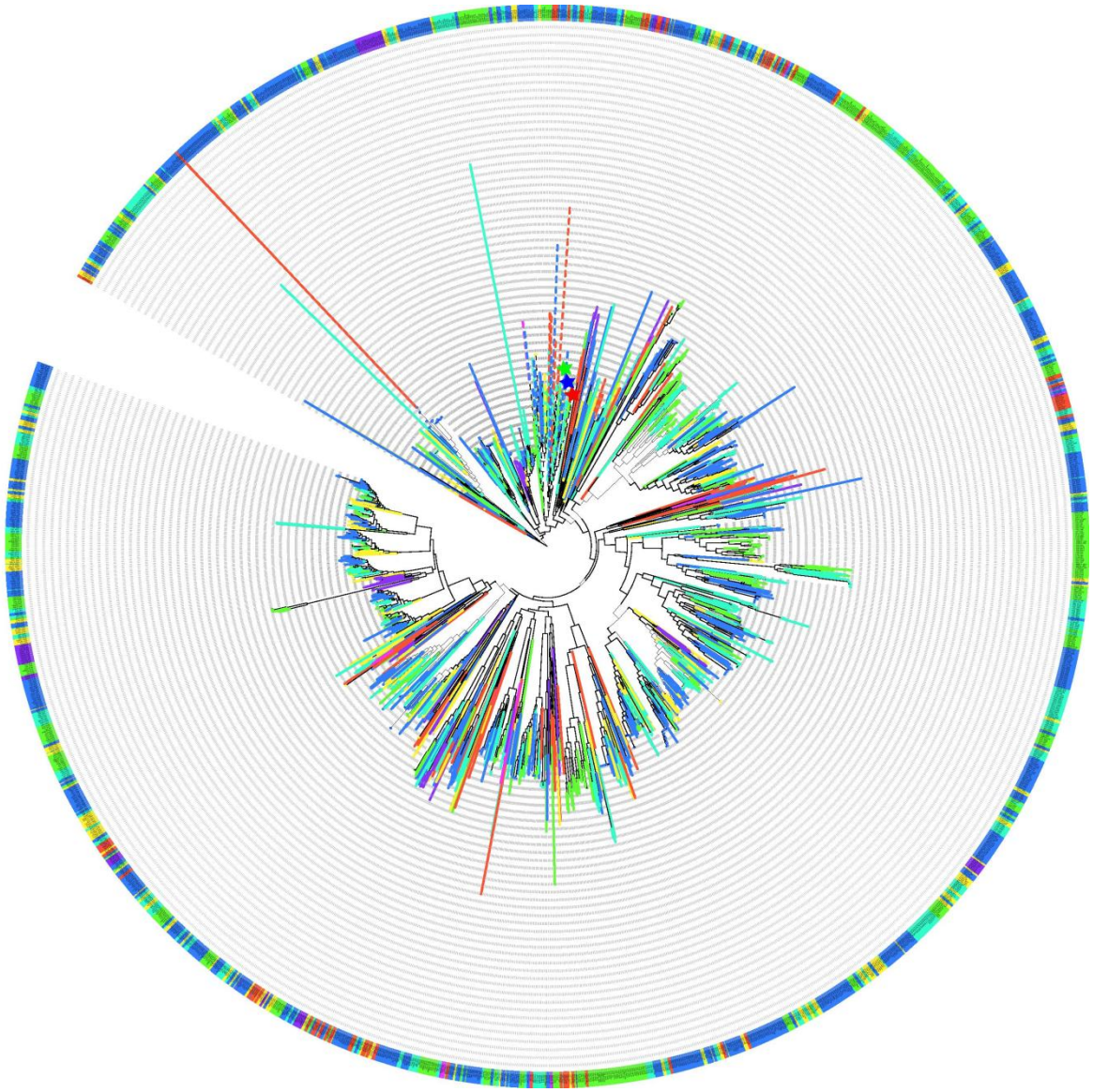


Figure 10. Phylogenetic tree of the Peroxisomal ATPase Pex1 reconstructed from the protein sequences that were detected by an HMM profile that was built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Selected peroxisomal clade is labeled with dashed lines.

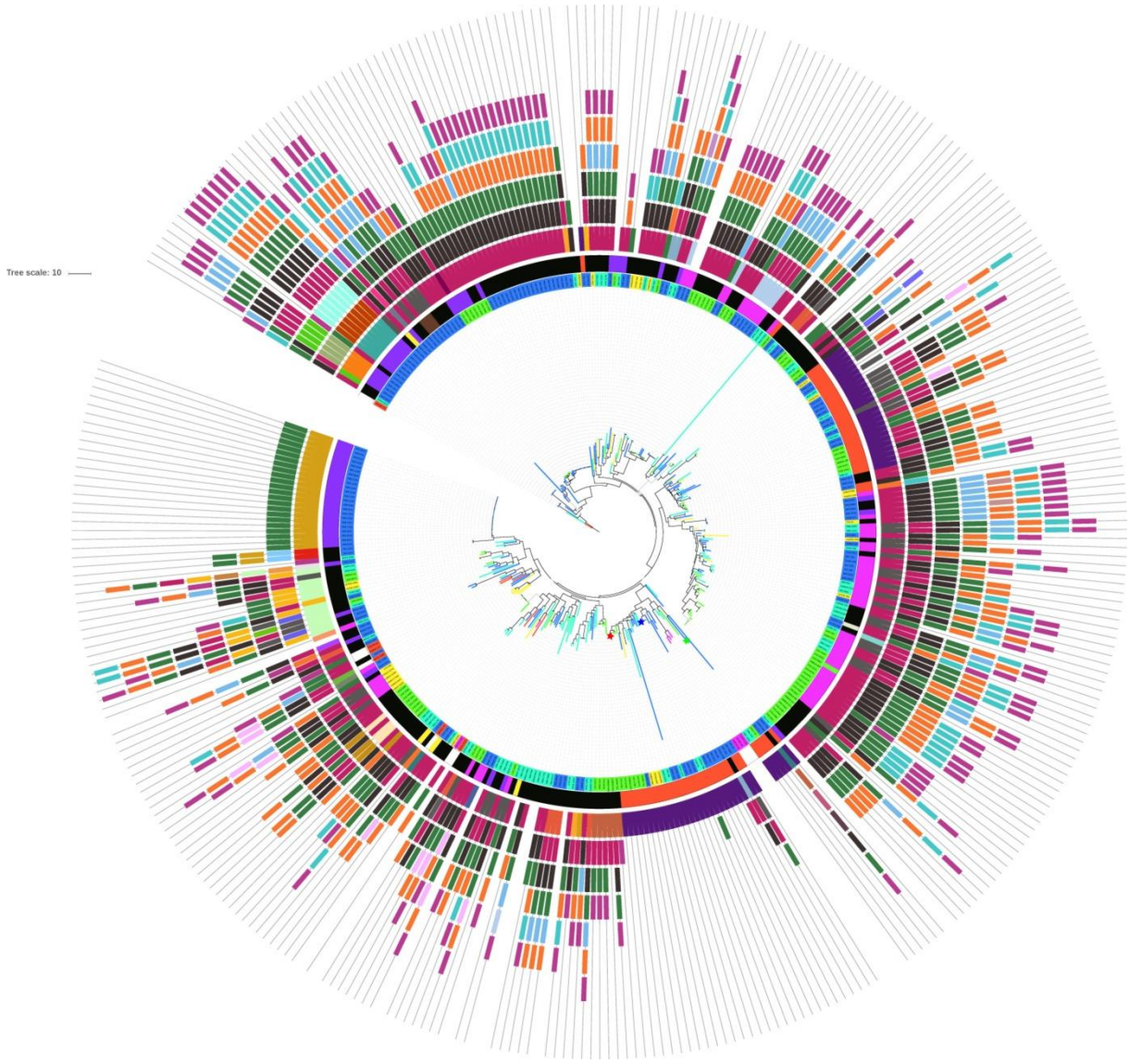


Figure 11. Phylogenetic tree of the Peroxisomal biogenesis factor 2 (Pex2) reconstructed from the protein sequences that were detected by HMM profile that was built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Tree was rooted using an outgroup method. In total 52 domains were found and the most common are Zinc finger C3HC4 type 1, Zinc finger C3HC4 type 2, Zinc finger C3HC4 type 3, Zinc finger C3HC4 type 4, Ring finger domain, RING-type zinc finger, zinc-RING finger domain and Pex2/Pex12 amino terminal region are shown in dark pink, dark gray, pine green, denim blue, rose, orange, plum, light green and the darkest purple.

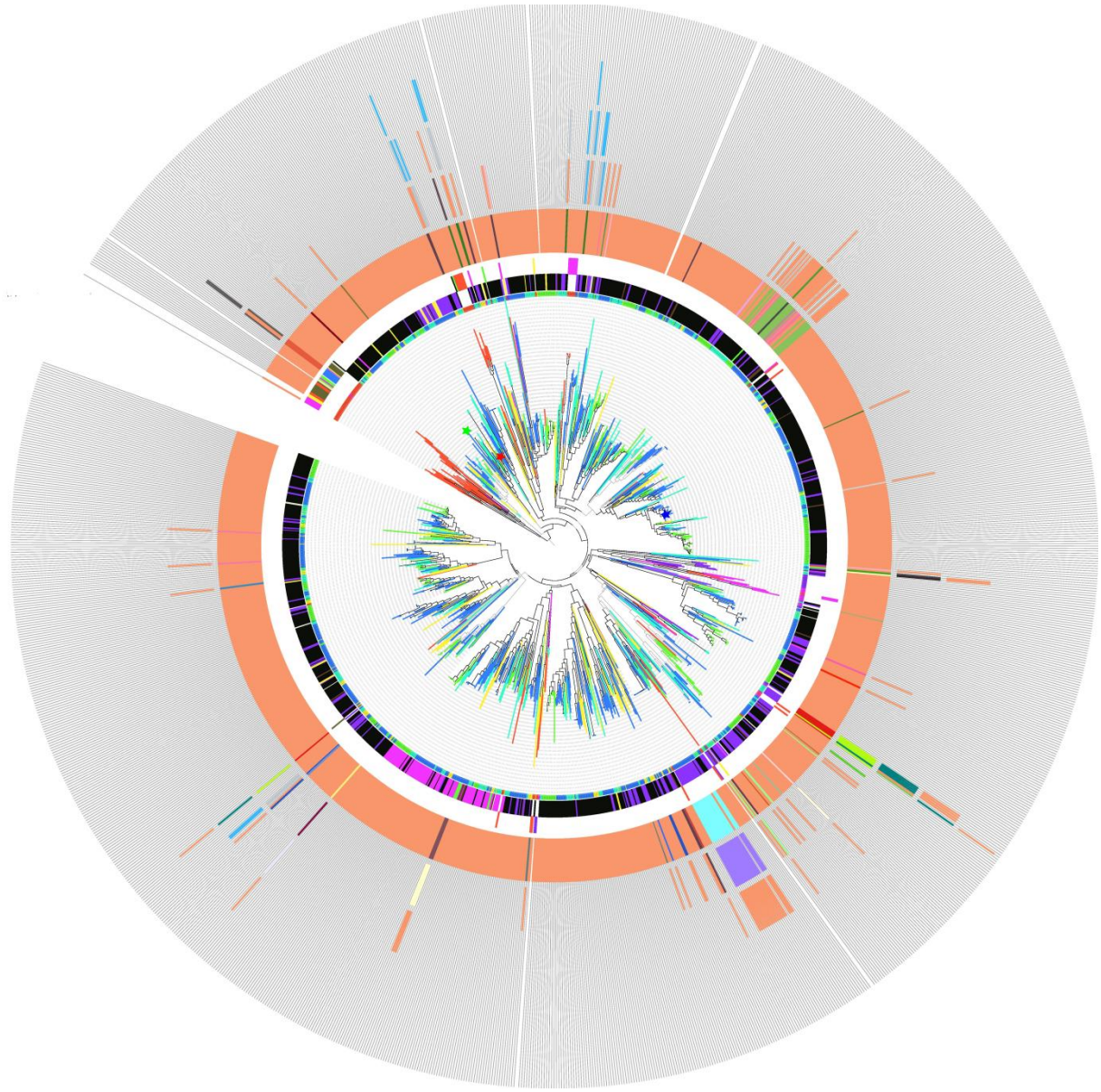


Figure 12. Phylogenetic tree of the Ubiquitin-conjugating enzyme E2 (Pex4) reconstructed from the protein sequences that were detected by HMM profiles that were built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Tree was rooted using an outgroup method. In total 50 different domains were found and the most common domain is Ubiquitin conjugating enzyme shown in peach color.

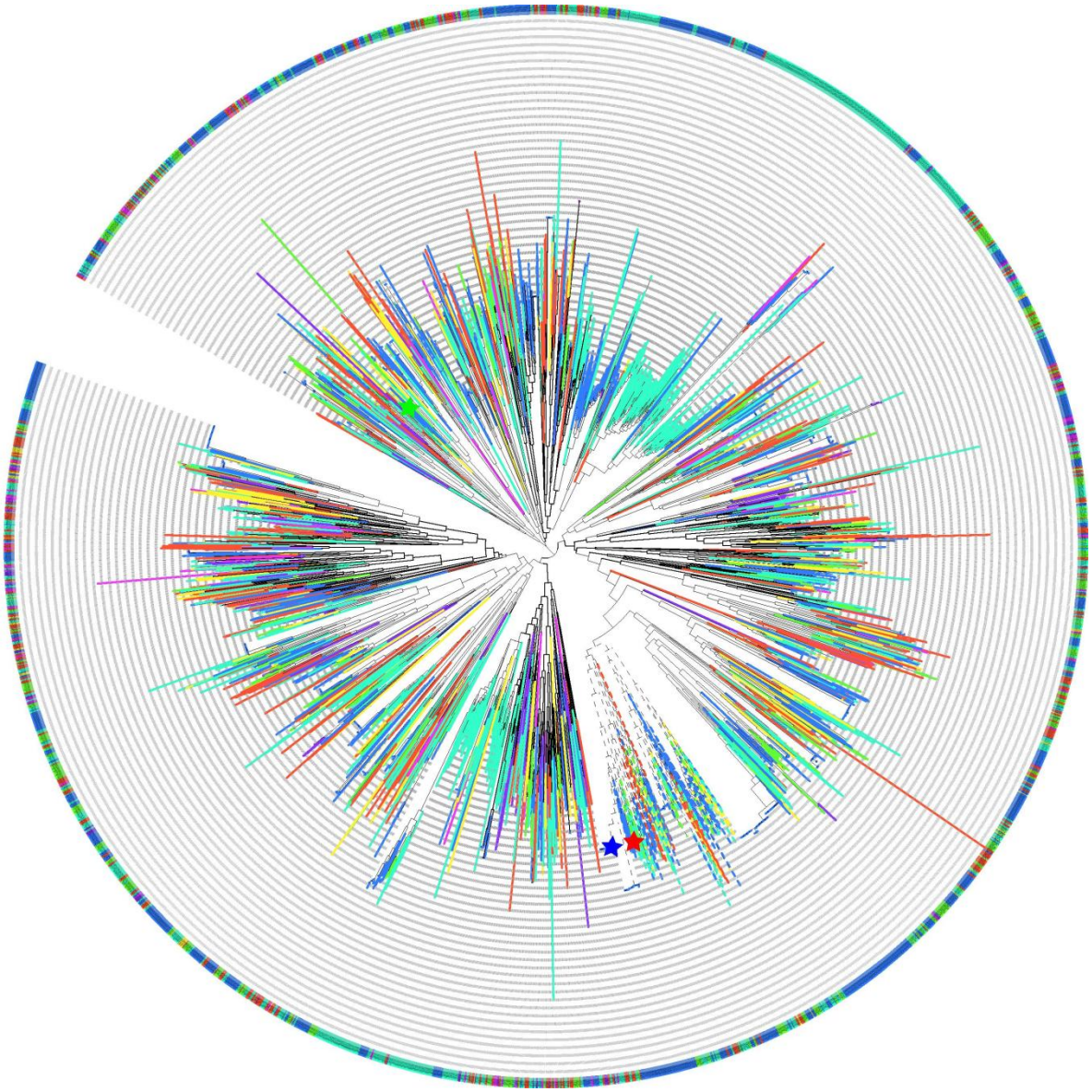


Figure 13. Phylogenetic tree of the Peroxisome biogenesis factor (Pex5) reconstructed from the protein sequences that were detected by an HMM profile that was built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Selected peroxisomal clade is labeled with dashed lines.

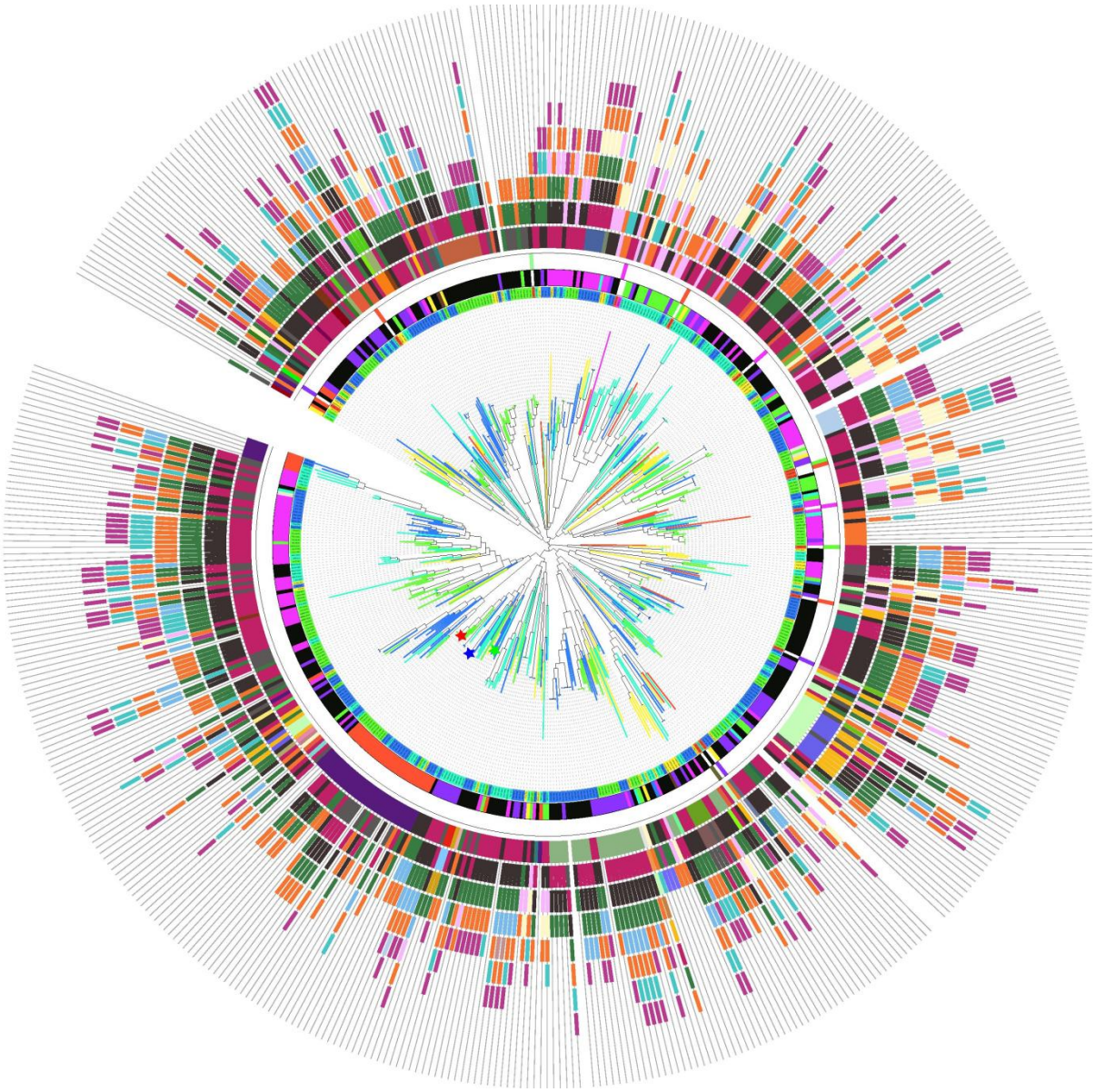


Figure 14. Phylogenetic tree of the Peroxisome biogenesis factor 10 (Pex10) reconstructed from the protein sequences that were detected by HMM profiles that were built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Tree was rooted using a midpoint method. In total 65 domains were detected and the most common one are Zinc finger C3HC4 type 1, Zinc finger C3HC4 type 2, Zinc finger C3HC4 type 3, Zinc finger C3HC4 type 4, Ring finger domain, RING-type zinc finger, zinc-RING finger domain, Pex2/Pex12 amino terminal region, RING-H2 zinc finger domain, RING-like zinc finger and SNF2 family N-terminal domain are shown dark pink, dark gray, pine green, denim blue, rose, orange, plum, light green, the darkest purple, blossom pink and lemon chiffon.

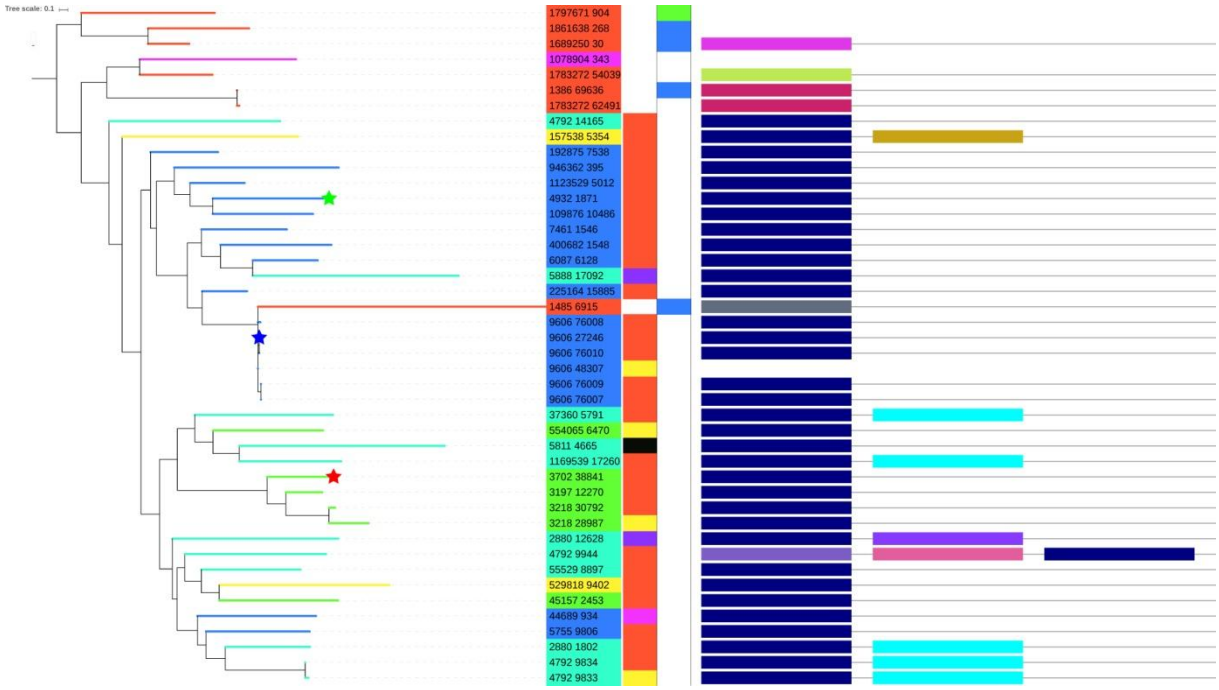


Figure 15. Phylogenetic tree of the Peroxisomal membrane protein 14 (Pex14) reconstructed from the protein sequences that were detected by HMM profiles that were built from the homologous sequences found in the BLAST search. Query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. Tree was rooted using an outgroup method. In total 10 domains are detected and the most common are Peroxisomal membrane anchor protein (Pex14p) conserved region and PUB domain shown in navy blue and cyan respectively.

3.4.2. Peroxisomal trees

Peroxisomal trees were derived from eukaryotic trees in a way that eukaryotic protein sequences from the peroxisomal sub tree were used to build an HMM profile with whom the prokaryotic part of the database was searched and 200 results with the highest e-value were taken as the closest prokaryotic relatives. Those 200 prokaryotic protein sequences were added to the set of eukaryotic sequences from the peroxisomal sub-tree and that newly formed set of sequences was used to reconstruct a phylogenetic tree using a phylogenetic pipeline described in Materials & Methods. Peroxisomal trees were built for proteins Pxa1, Pxa2, Faa2, Pex1 and Pex5.

Table 10. For each protein for whom a peroxisomal tree was reconstructed total number of protein sequences and its distribution across Bacteria, Archaea, Asgard group and Eukaryota is shown.

Protein	Number of hits	Eukaryotes	Bacteriae	Archaeae	Asgard
Pxa1	340	140	200	0	0
Pxa2	351	151	200	0	0
Faa2	531	331	198	2	0
Pex1	283	83	94	84	22
Pex5	437	237	150	50	0

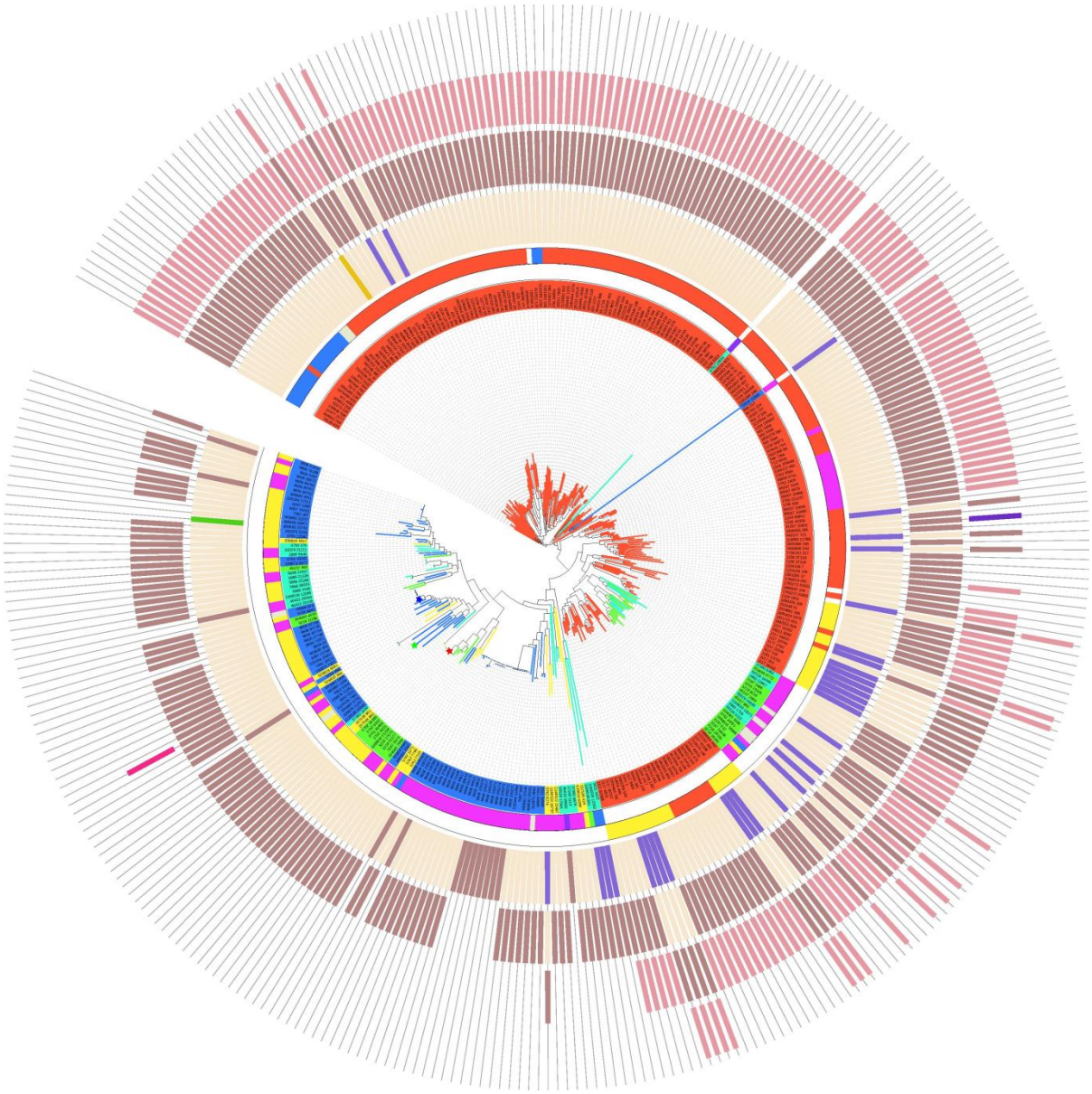


Figure 16. Peroxisomal tree of the protein Peroxisomal long-chain fatty acid import protein 2 (Pxa1) reconstructed from the eukaryotic sequences of the peroxisomal sub tree and 200 prokaryotic sequences with the highest e-value in the HMM search performed with the profile built from the eukaryotic sequences on the prokaryotic part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 8 different domains were detected and the most common are ABC transporter domain, ABC transporter transmembrane region 2, SbmA/BacA-like family and ABC transporter transmembrane region shown in light purple, beige, pink and purple respectively. Clade of bacterial sequences from phyla Firmicutes was used as outgroup.

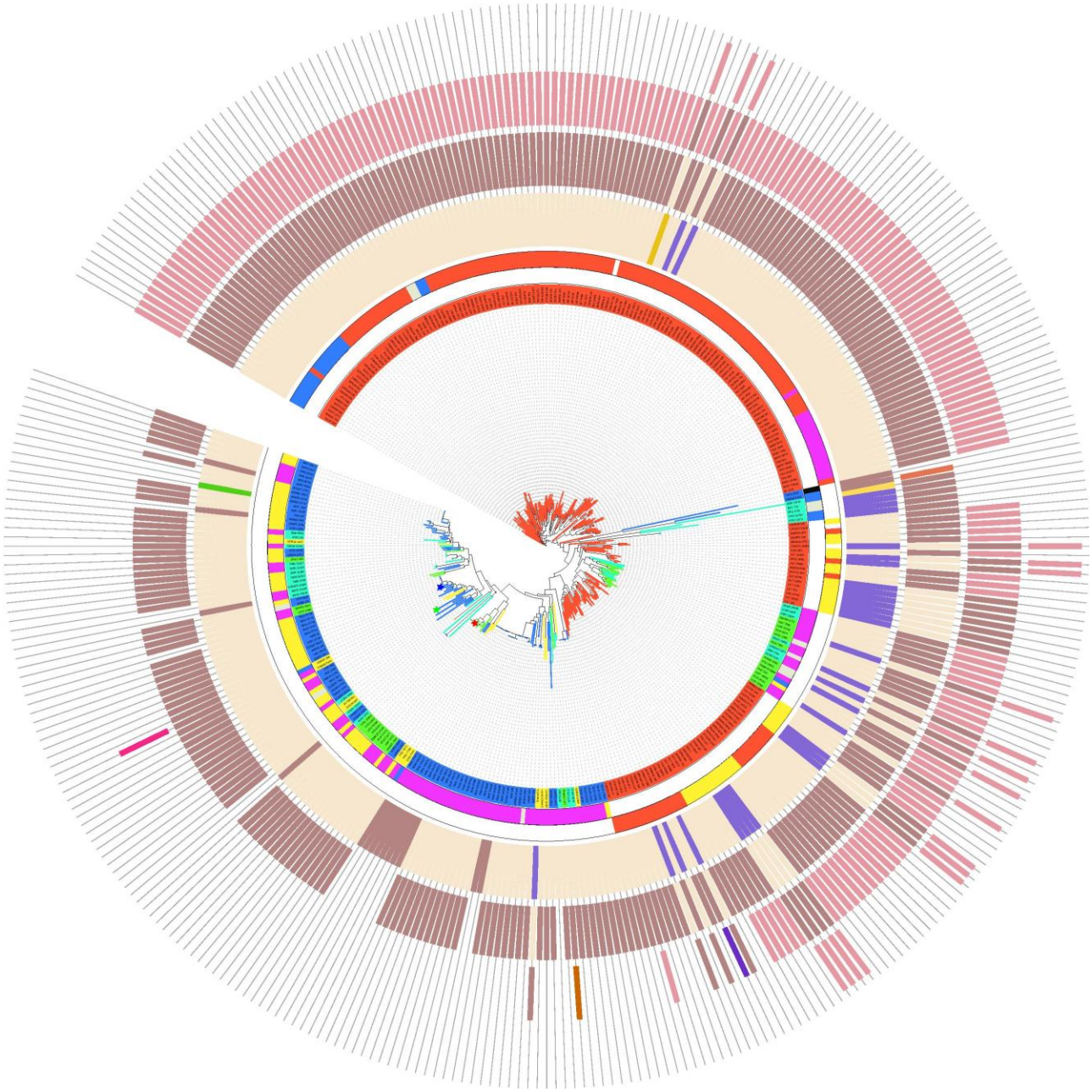


Figure 17. Peroxisomal tree of the protein Peroxisomal long-chain fatty acid import protein 1 (Pxa2) reconstructed from the eukaryotic sequences of the peroxisomal sub tree and 200 prokaryotic sequences with the highest e-value in the HMM search performed with the profile built from the eukaryotic sequences on the prokaryotic part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 11 different domains were detected and the most common are ABC transporter domain, ABC transporter transmembrane region 2, SbmA/BacA-like family and ABC transporter transmembrane region shown in light purple, beige, pink and purple respectively. Clade of bacterial sequences from phyla Firmicutes was used as outgroup.

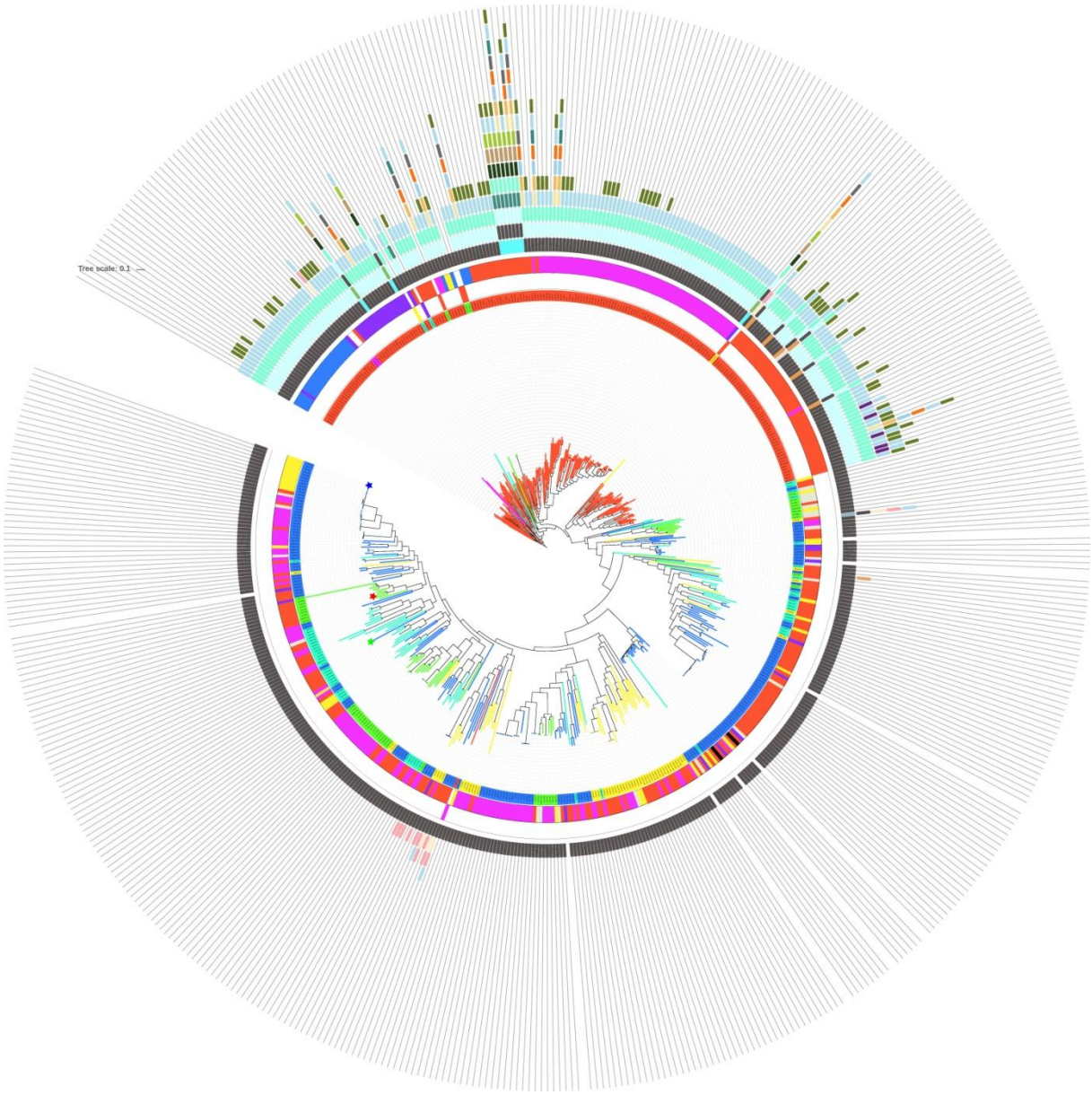


Figure 18. Peroxisomal tree of the protein Long-chain-fatty-acid-CoA ligase 2 (Faa2) reconstructed from the eukaryotic sequences of the peroxisomal sub tree and 200 prokaryotic sequences with the highest e-value in the HMM search performed with the profile built from the eukaryotic sequences on the prokaryotic part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 24 domains were found and the most common are AMP-binding enzyme domain, Phosphopantetheine attachment site, AMP-binding enzyme C-terminal domain, Condensation domain and Thioesterase domain shown in gray, light blue, the lightest blue, aquamarine and fern green. Clade of bacterial sequences from phyla Firmicutes was used as outgroup.

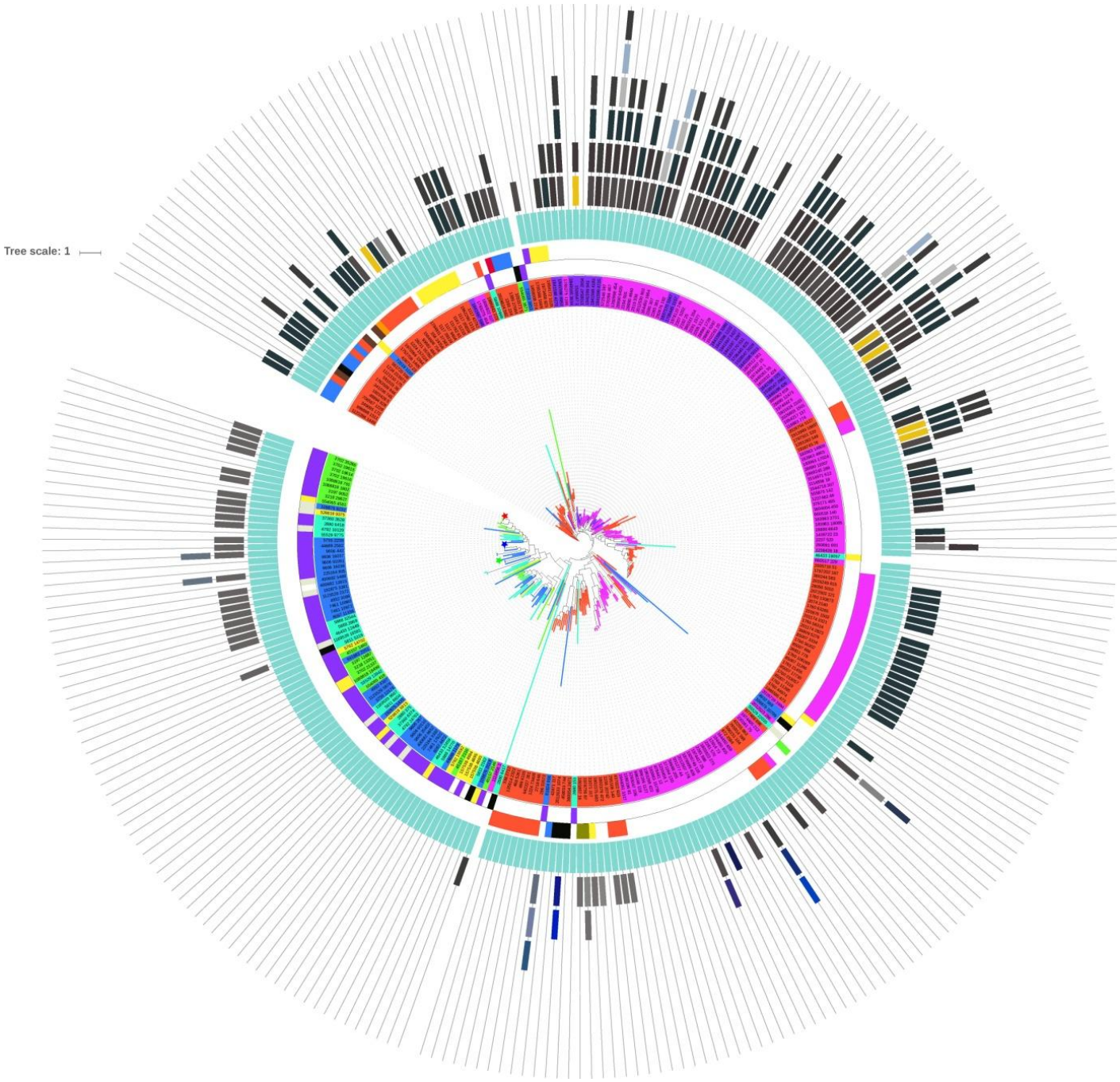


Figure 19. Peroxisomal tree of the protein Peroxisomal biogenesis factor 1 (Pex1) reconstructed from the eukaryotic sequences of the peroxisomal sub tree and 200 prokaryotic sequences with the highest e-value in the HMM search performed with the profile built from the eukaryotic sequences on the prokaryotic part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 22 domains were found and the most common are ATPase family associated with various cellular activities (AAA), Cell division protein 48 (CDC48) domain 2, Cell division protein 48 (CDC48) N-terminal domain, Holiday junction DNA helicase ruvB N-terminus, AAA domain (CDC48 subfamily) and Peroxisome biogenesis factor 1 N-terminal domain are shown in light blue, the darkest gray, dark slate gray, iridium, gray wolf and ash gray. Clade of bacterial sequences from phyla Firmicutes was used as outgroup.

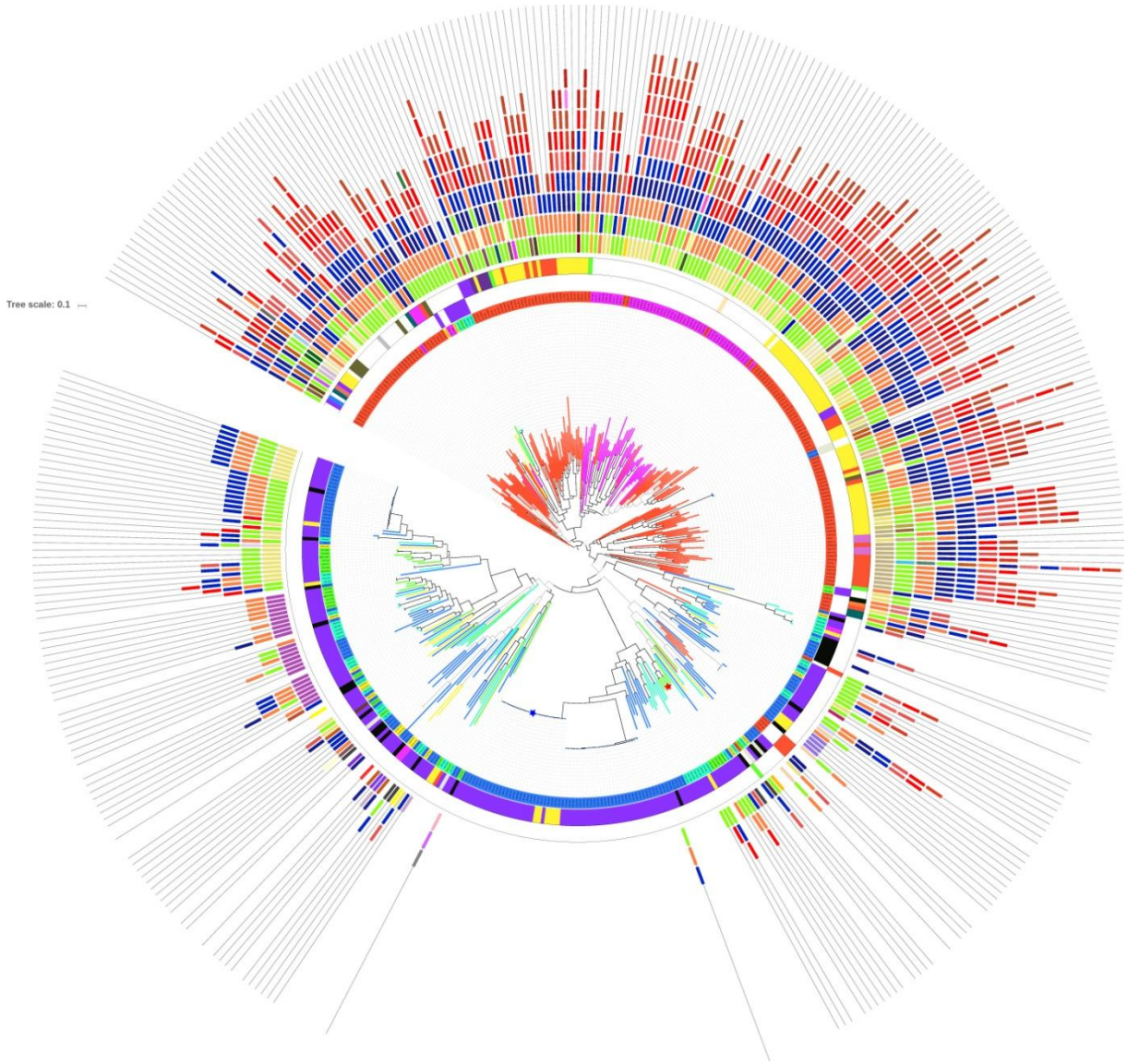


Figure 20. Peroxisomal tree of the protein Peroxisomal biogenesis factor 5 (Pex5) reconstructed from the eukaryotic sequences of the peroxisomal sub tree and 200 prokaryotic sequences with the highest e-value in the HMM search performed with the profile built from the eukaryotic sequences on the prokaryotic part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 22 domains were found and the most common are Tetratricopeptide repeat 1, Tetratricopeptide repeat 11, Tetratricopeptide repeat 16, Tetratricopeptide repeat 12, Tetratricopeptide repeat 2, Tetratricopeptide repeat 19, Tetratricopeptide repeat 8, Tetratricopeptide repeat 9, Anaphase-promoting complex subunit 3, Tetratricopeptide repeat 17 and Anaphase-promoting complex subunit 8 are shown in chartreuse, orange, cobalt blue, dark blue, red, valentine red, grapefruit, chestnut red, harvest gold, bean red and medium orchid. Tree was rooted using an outgroup method.

3.4.3. Phylogenetic trees based on orthology search

Results were ordered according to the e-value and the 200th prokaryotic hit was a cut-off value for a number of sequences that will be used for reconstruction of a phylogenetic tree. In Table 11. total number of hits, selected number of hits and their distribution across domains Eukaryota, Bacteria, Archaea and Asgard group. For proteins Pex2, Pex4 and Pex10 that had less than 200 prokaryotic hits first 500 results were selected while for Pex14 all proteins were taken since their number was low enough to build a phylogenetic tree using a reliable phylogenetic pipeline.

All trees from this set were reconstructed using a phylogenetic pipeline that is described in Materials & Methods section besides for protein Faa2 for whom MUSCLE alignment couldn't be obtained so MAFFT and –auto parameter was used for the alignment and Fasttree with –wag parameter was used for the reconstruction of the phylogenetic tree. Trees are visualized in the same way as eukaryotic and peroxisomal trees.

Table 11. For each protein total number of hits found are found with HMM profiles built from the orthologous protein sequences is shown together with number of selected proteins and their distribution across domains Eukaryota, Bacteria, Archaea and Asgard group.

Protein	Total	Top 200	Eukaryotes	Bacteriae	Archaeae	Asgard
Cta1	570	270	70	191	5	4
Pxa1	94,646	504	304	200	0	0
Pxa2	54,379	339	139	200	0	0
Fox1	2,624	367	167	200	0	0
Fox2	51,501	240	40	196	3	1
Faa2	36,946	401	201	197	3	0
Pex1	9,265	459	259	93	84	23
Pex2	1,237	500	491	7	2	0
Pex4	1175	500	490	8	2	0
Pex5	38,954	292	92	57	43	0
Pex10	4,388	500	494	6	0	0
Pex14	146	146	139	7	0	0

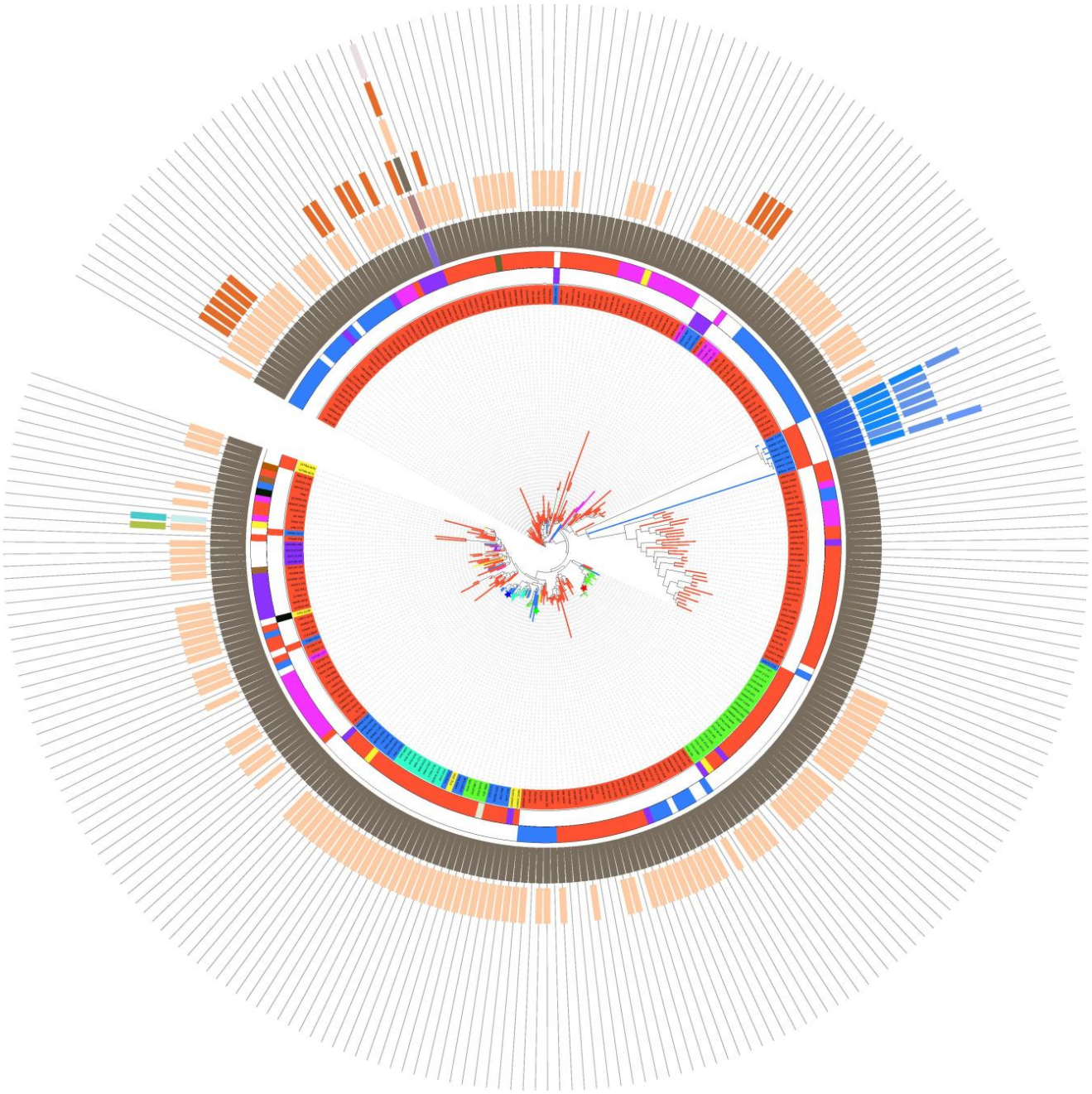


Figure 21. Phylogenetic tree of the protein Catalase (Cta1) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 14 different domains were detected and two most common ones are catalase domain shown in grey and catalase-related immune responsive domain shown in beige color that are found in eukaryotic, bacterial and archael protein sequences. Clade of bacterial sequences from phyla Firmicutes was used as outgroup.

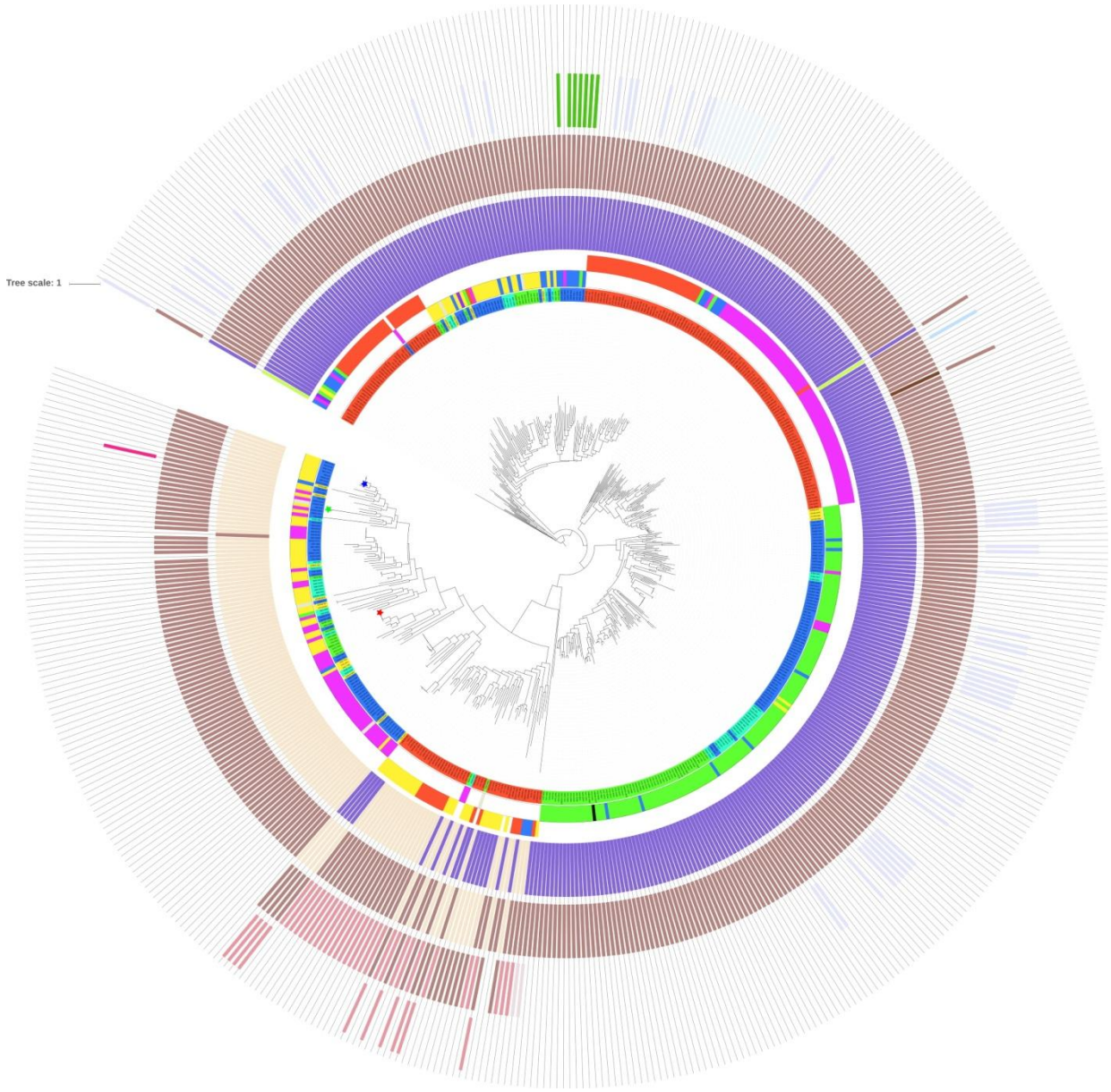


Figure 22. Phylogenetic tree of the protein Peroxisomal long-chain fatty acid import protein 2 (Pxa1) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 12 different domains were detected and the most common are ABC transporter domain, ABC transporter transmembrane region 2, SbmA/BacA-like family, ABC transporter transmembrane region and RecF/RecN/SMC N terminal domain shown in light purple, beige, pink, purple and pale blue respectively. Bacterial sequence from phyla Firmicutes was used as outgroup.

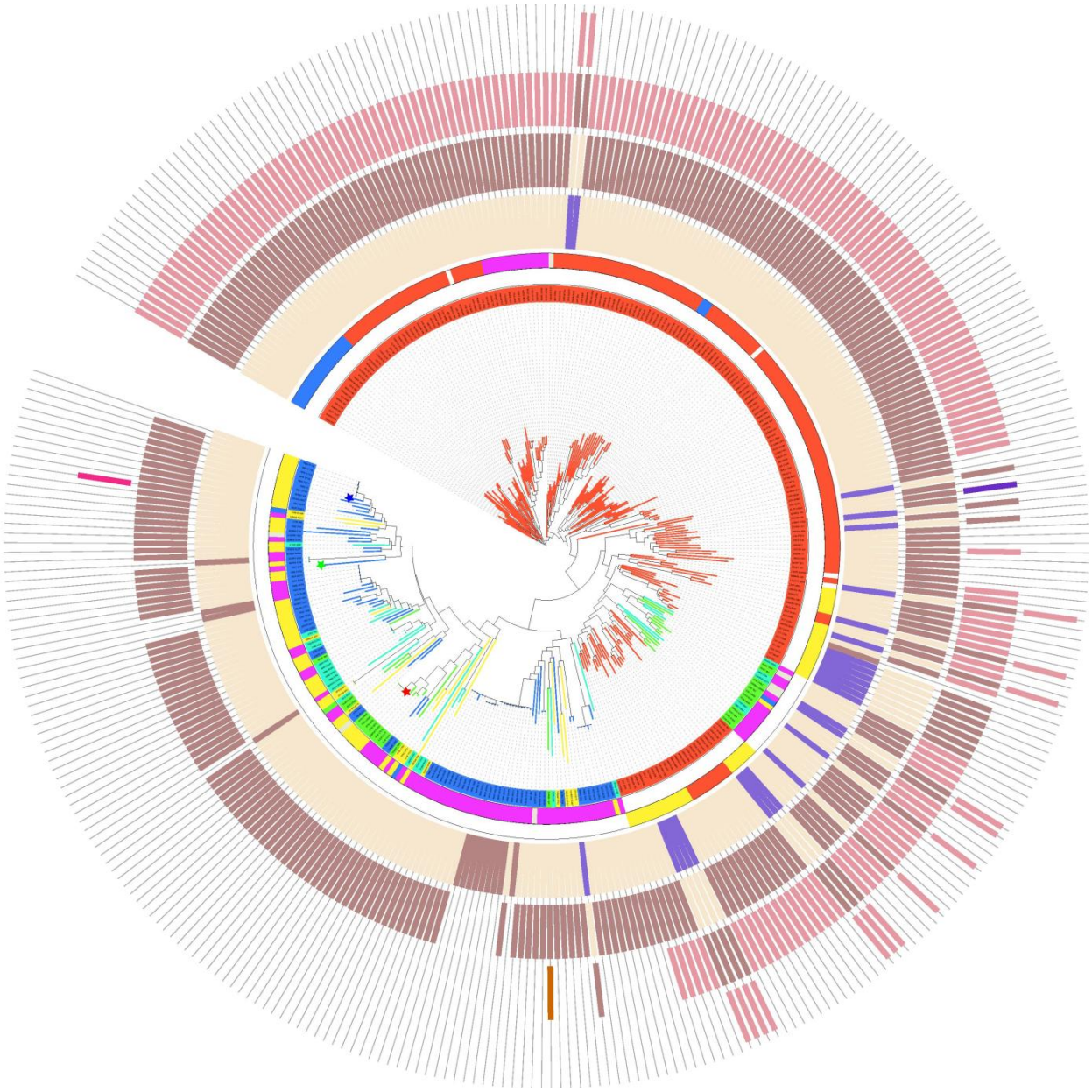


Figure 23. Phylogenetic tree of the protein Peroxisomal long-chain fatty acid import protein 1 (Pxa2) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 7 different domains were detected and the most common are ABC transporter domain, ABC transporter transmembrane region 2, SbmA/BacA-like family and ABC transporter transmembrane region shown in light purple, beige, pink and purple respectively. Clade of bacterial sequences from phyla Firmicutes was used as outgroup.

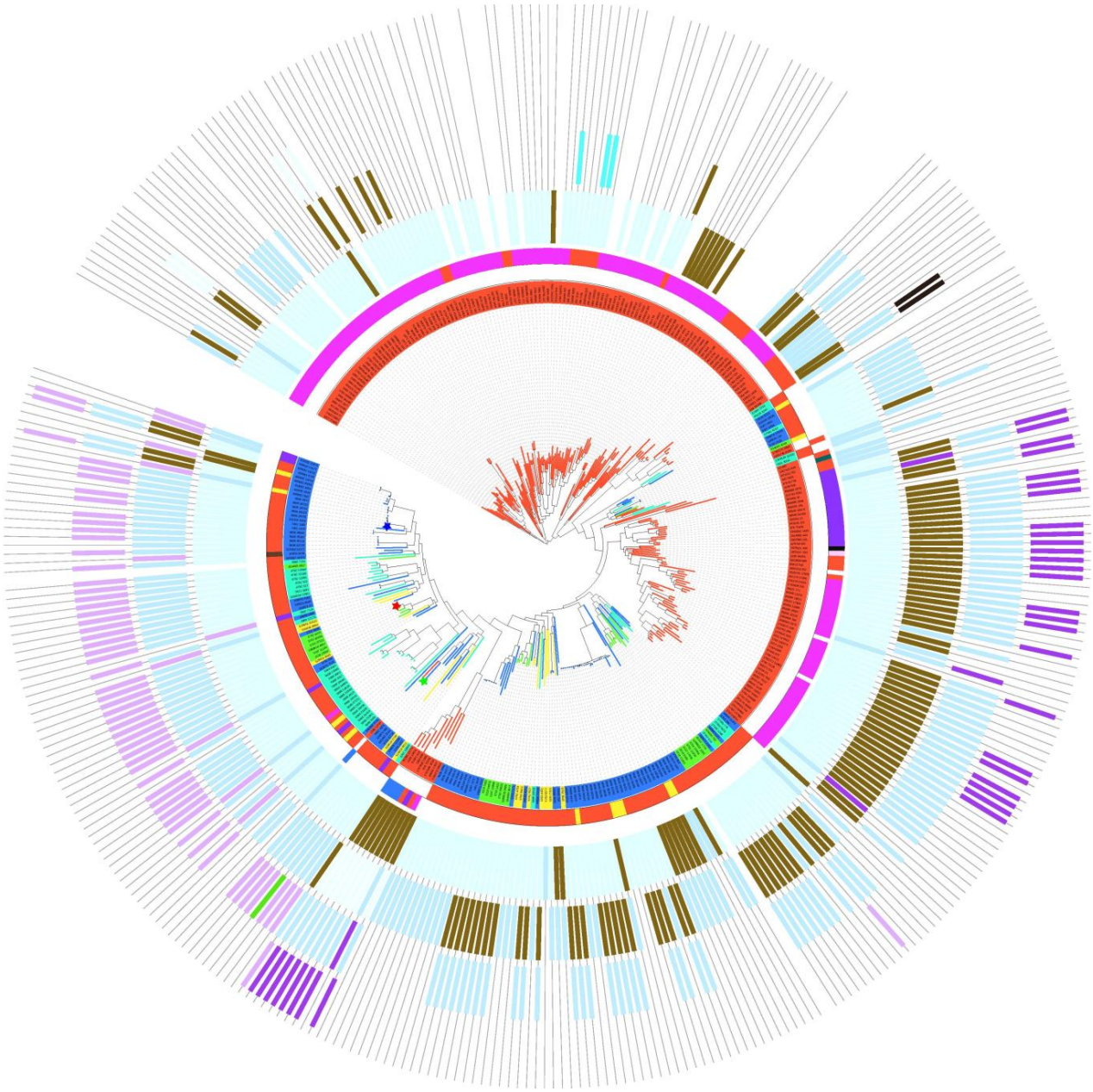


Figure 24. Phylogenetic tree of the protein Acyl-coenzyme A oxidase (Fox1) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 9 domains were detected and the most common are Acyl-CoA dehydrogenase middle domain, Acyl-CoA dehydrogenase C-terminal domain 1, Acyl-CoA dehydrogenase N-terminal domain, Acyl-CoA dehydrogenase C-terminal domain 2, Acyl-CoA oxidase domain and Acyl-CoA oxidase N-terminal domain are shown in light blue, oak brown, purple, azure, light cyan and mauve respectively.

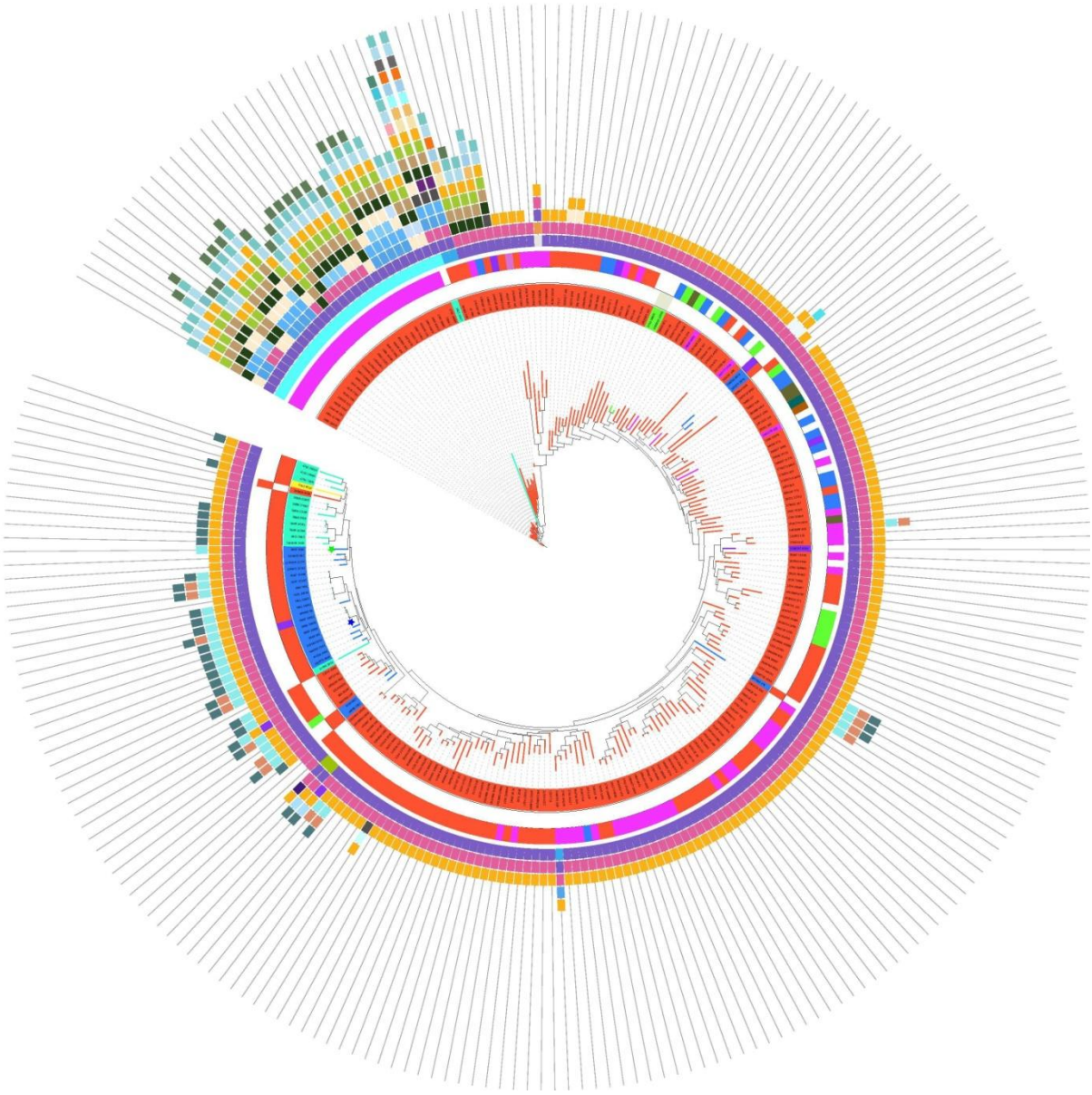


Figure 25. Phylogenetic tree of the protein Peroxisomal hydratase-dehydrogenase-epimerase (Fox2) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 40 domains are detected and the most common are Short chain dehydrogenase, KR domain, Enoyl-(Acyl carrier protein) reductase, Ketoacyl-synthetase C-terminal extension, Beta-ketoacyl synthase N-terminal domain, Beta-ketoacyl synthase C-terminal domain, Phosphopantetheine attachment domain, Polyketide synthase dehydratase, SCP2-sterol transfer family, MaoC like domain, Acyl transferase domain, and NAD-dependent epimerase/dehydratase family are shown in purple, golden yellow, pink, dark green, beige, green, light blue, light green, medium green, the lightest blue, neon blue and almond respectively.

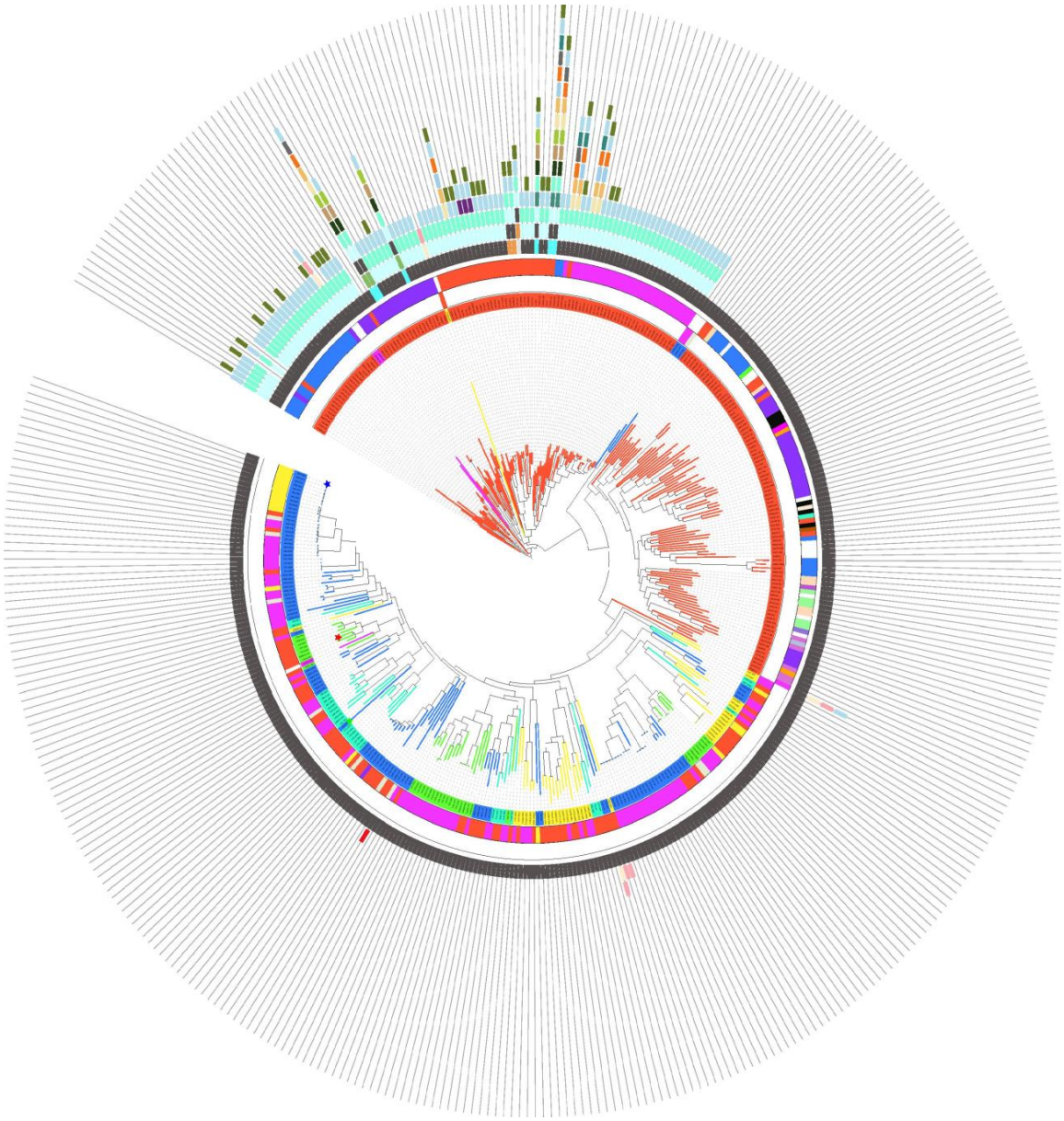


Figure 26. Phylogenetic tree of the protein Long-chain-fatty-acid—CoA ligase 2 (Faa2) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 23 domains were found and the most common are AMP-binding enzyme domain, Phosphopantetheine attachment site, AMP-binding enzyme C-terminal domain, Condensation domain and Thioesterase domain shown in gray, coral blue, light slate, aquamarine and fern green respectively.

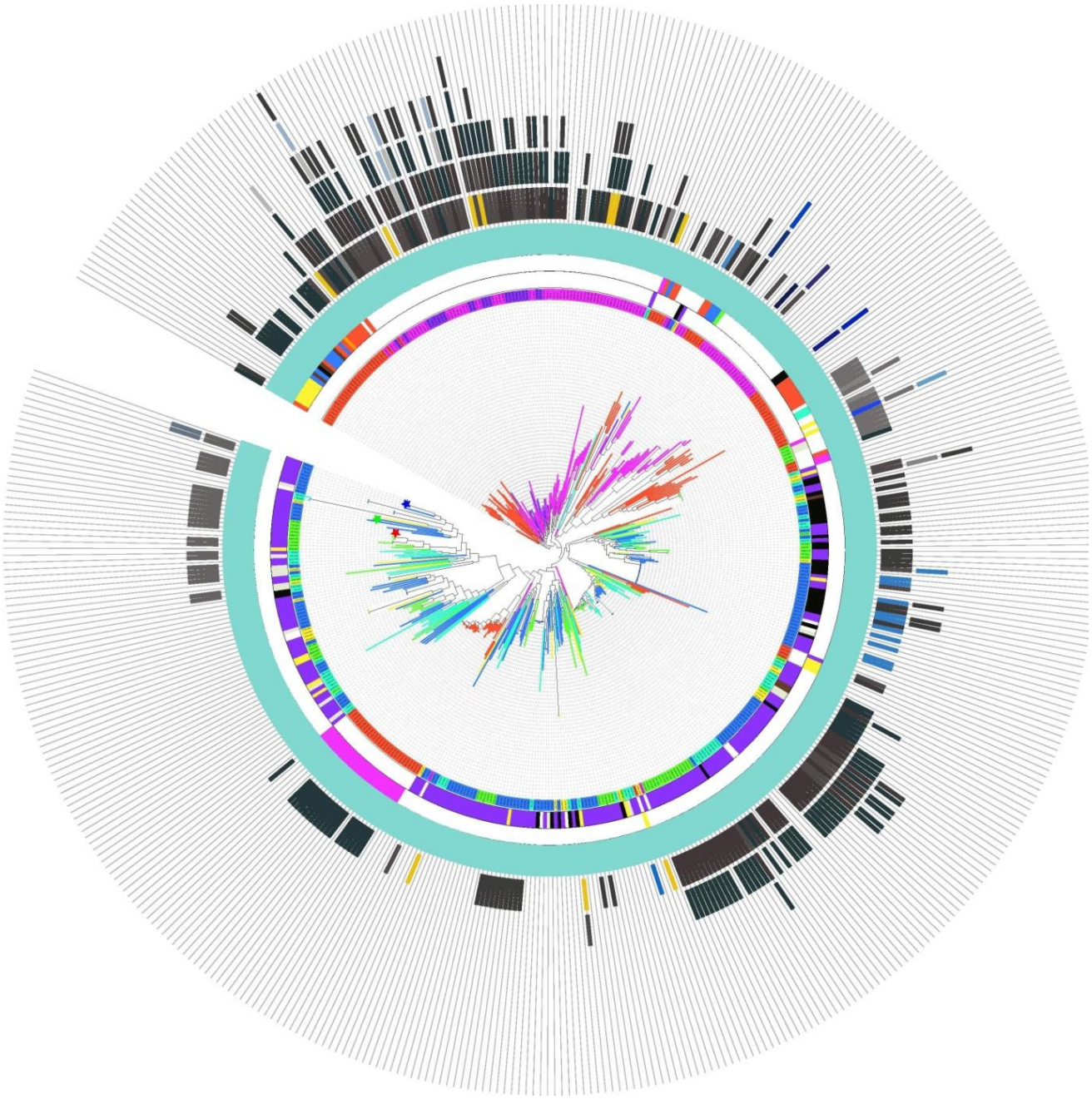


Figure 27. Phylogenetic tree of the protein Peroxisome biogenesis factor 1 (Pex1) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 23 different domains were detected and the most common are ATPase family associated with various cellular activities (AAA), Cell division protein 48 (CDC48) domain 2, Cell division protein 48 (CDC48) N-terminal domain, Holiday junction DNA helicase ruvB N-terminus, AAA domain (Cdc48 subfamily) and Peroxisome biogenesis factor 1 N-terminal domain shown in light blue, the darkest gray, dark slate gray, iridium, gray wolf and ash gray respectively.

.

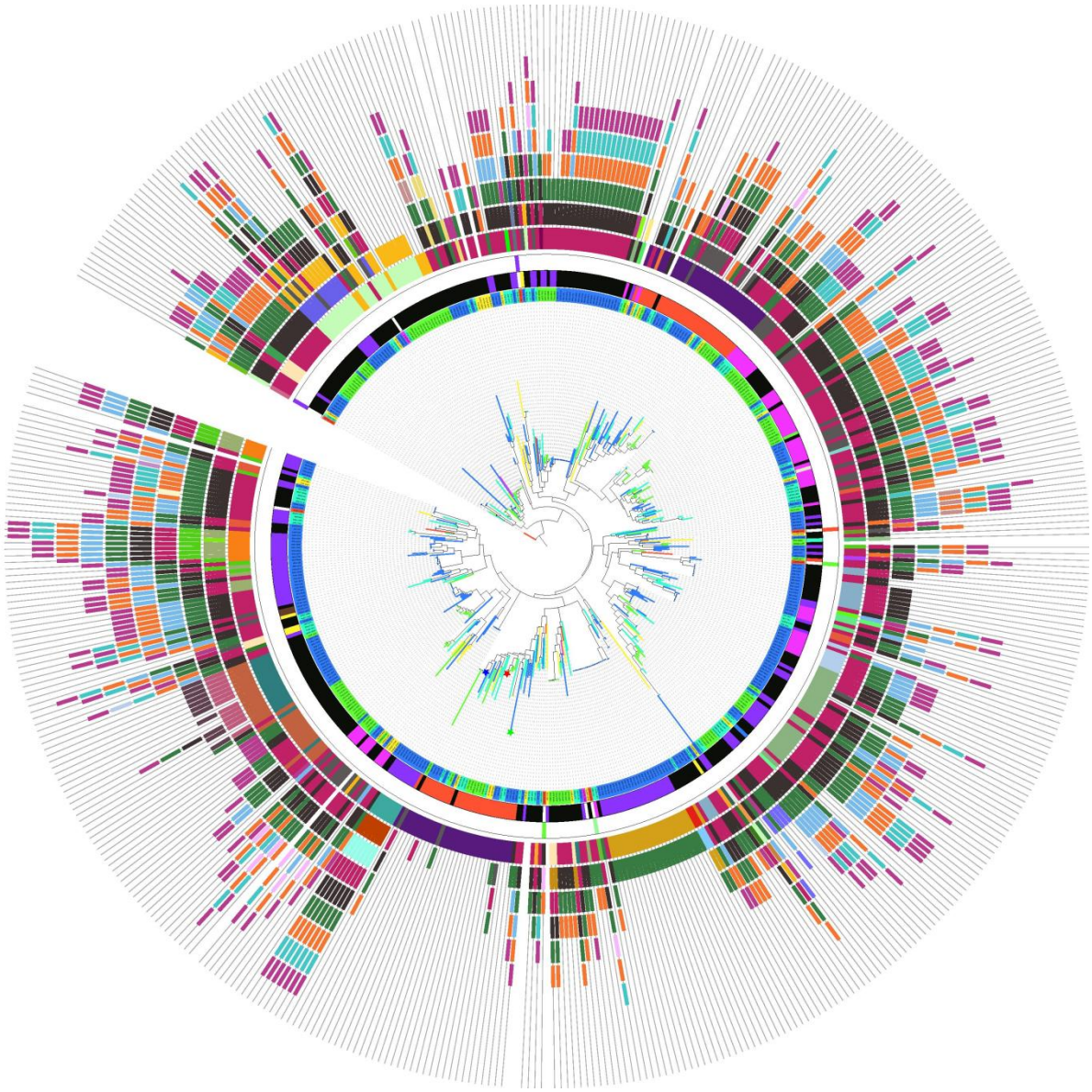


Figure 28. Phylogenetic tree of the protein Peroxisome biogenesis factor 2 (Pex2) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 65 domains were found and the most common are Zinc finger C3HC4 type 1, Zinc finger C3HC4 type 2, Zinc finger C3HC4 type 3, Zinc finger C3HC4 type 4, Ring finger domain, RING-type zinc finger, zinc-RING finger domain and Pex2/Pex12 amino terminal region are shown in dark pink, dark gray, pine green, denim blue, rose, orange, plum, light green and the darkest purple respectively.

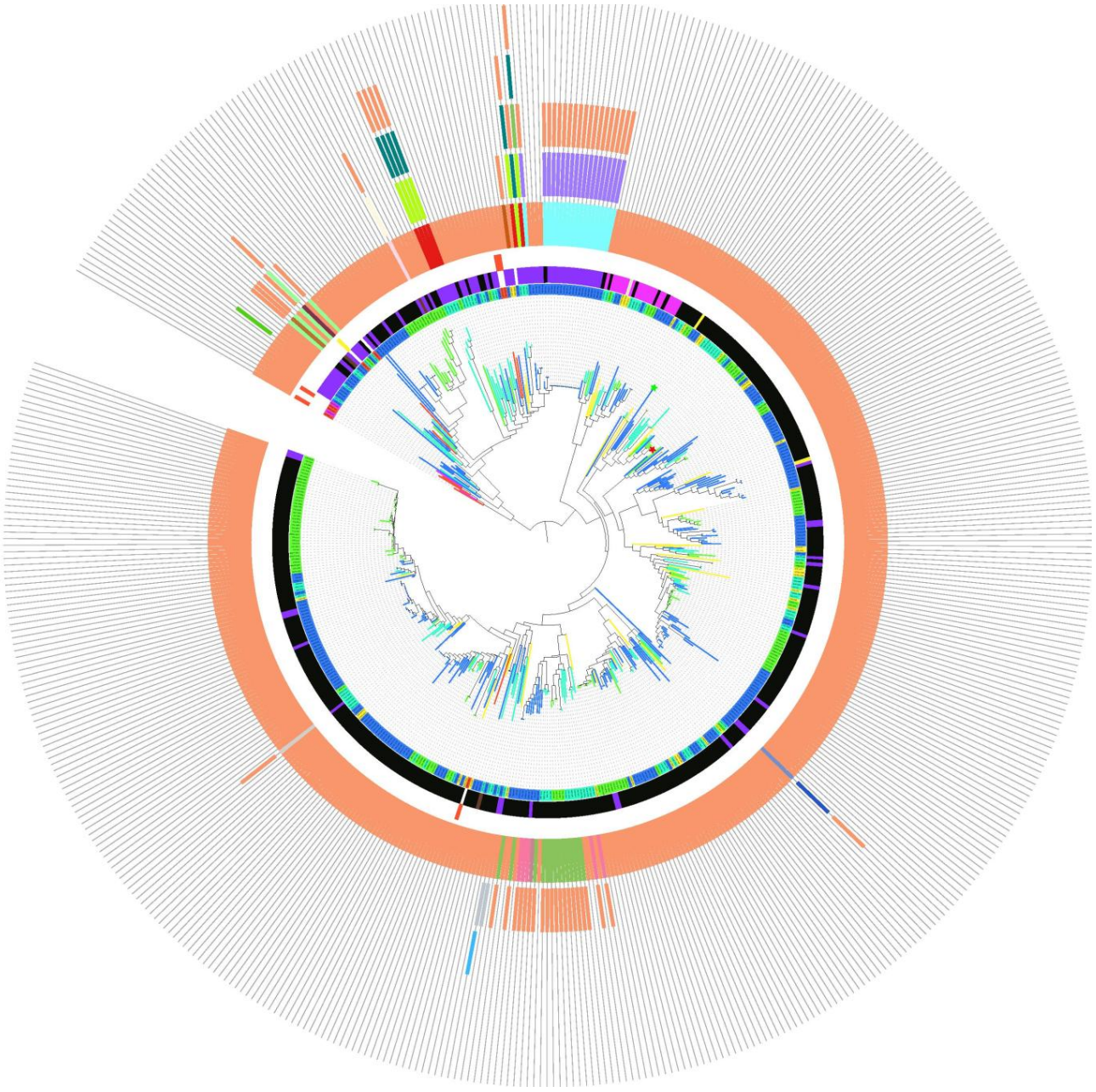


Figure 29. Phylogenetic tree of the protein Ubiquitin-conjugating enzyme E2-21 kDa (Pex4) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 21 different domains were found and the most common domain is Ubiquitin conjugating enzyme domain shown in peach color.

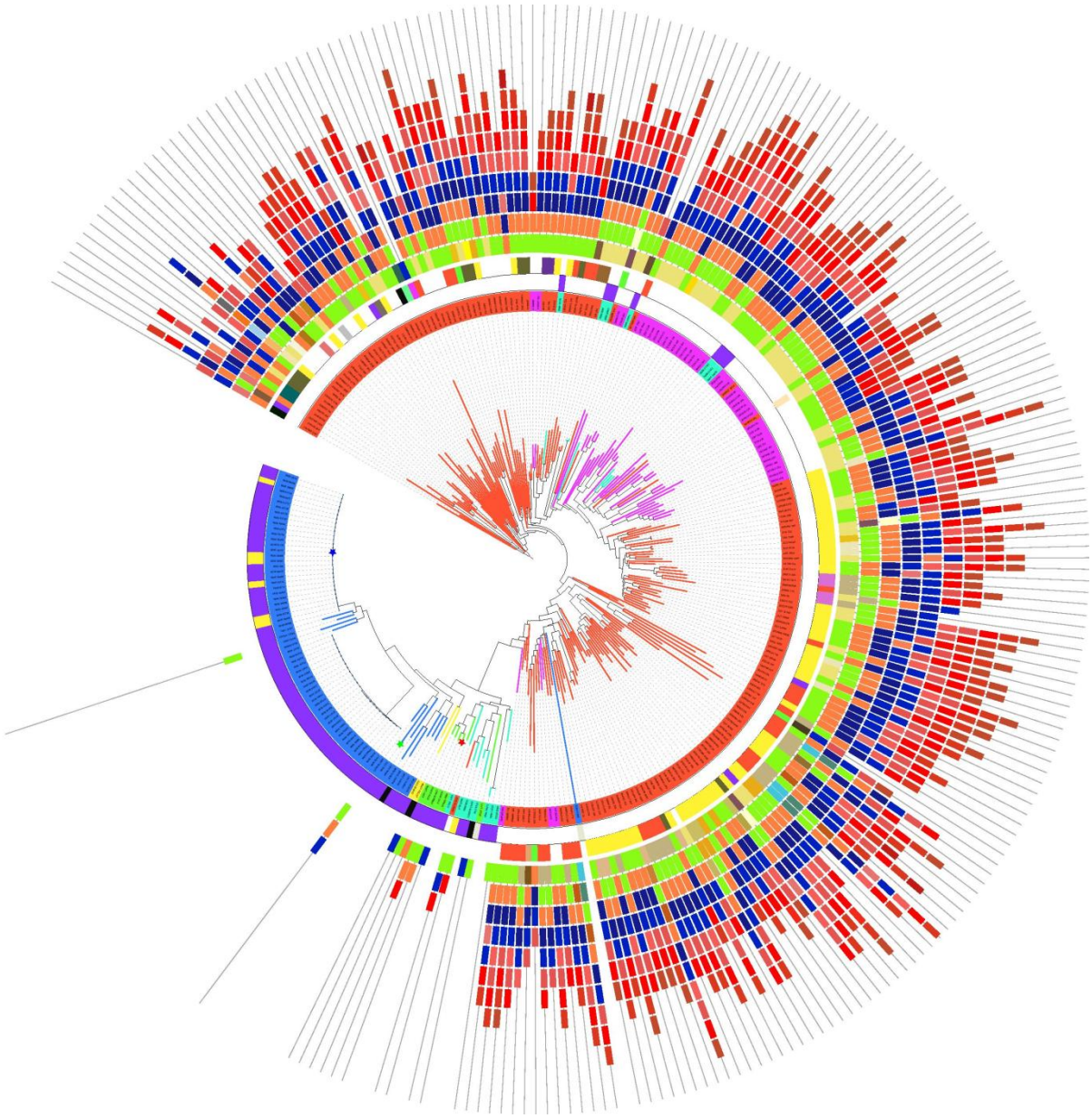


Figure 30. Phylogenetic tree of the protein Peroxisomal targeting signal 1 receptor(Pex5) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 47 domains were detected and the most common are Tetratricopeptide repeat 11, Tetratricopeptide repeat 16, Tetratricopeptide repeat 1, Tetratricopeptide repeat 12, Tetratricopeptide repeat 2, Tetratricopeptide repeat 19, Tetratricopeptide repeat 8, Tetratricopeptide repeat 9, Tetratricopeptide repeat 17 and Anaphase-promoting complex cyclosome subunit 3 shown in orange, cobalt blue, light green, denim dark blue, red, dark red, grapefruit, Chestnut red, bean red and harvest gold respectively.

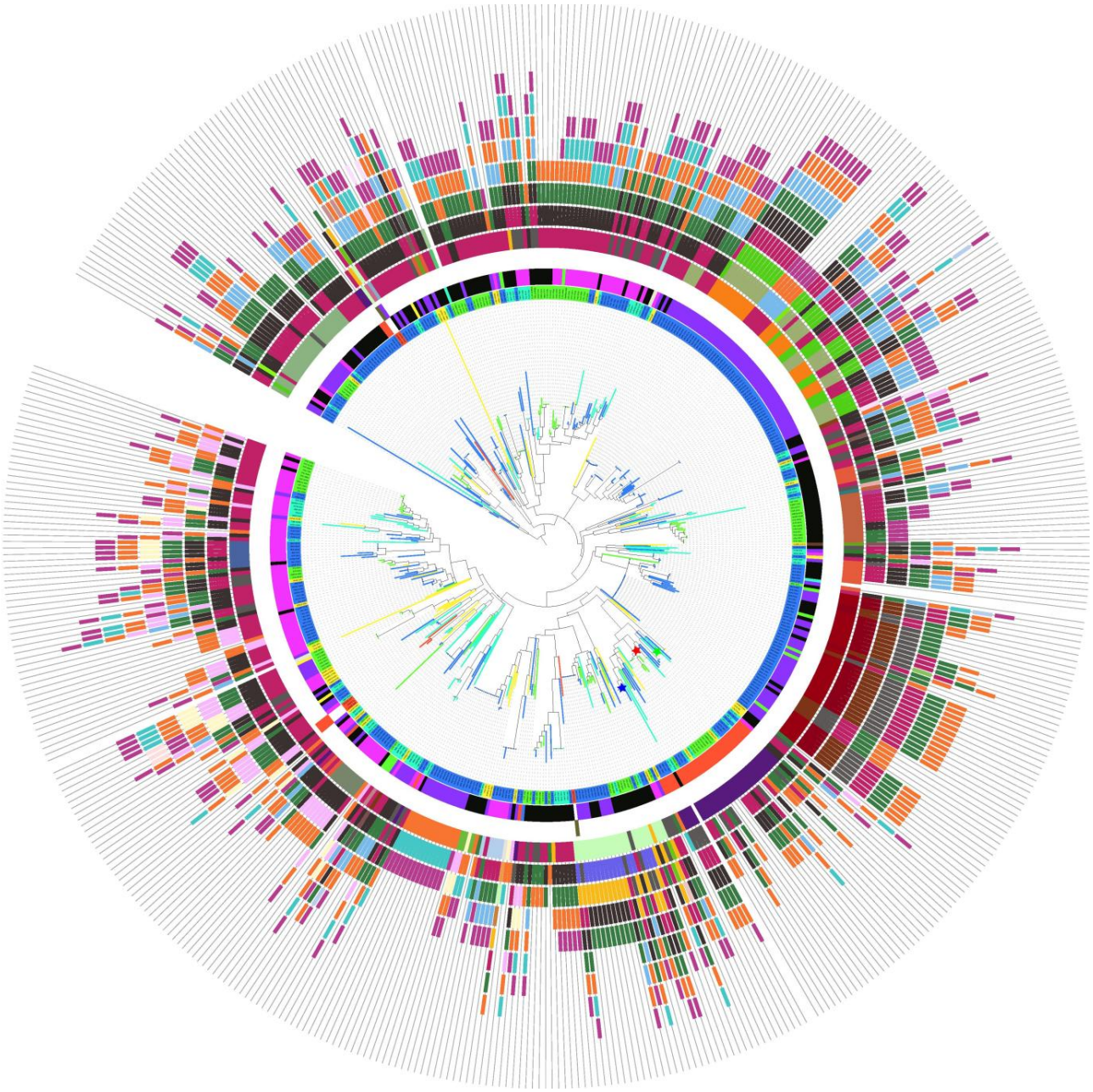


Figure 31. Phylogenetic tree of the protein Peroxisome biogenesis factor (Pex10) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 55 domains were detected and the most common one are Zinc finger C3HC4 type 1, Zinc finger C3HC4 type 2, Zinc finger C3HC4 type 3, Zinc finger C3HC4 type 4, Ring finger domain, RING-type zinc finger, zinc-RING finger domain, Pex2/Pex12 amino terminal region, RING-H2 zinc finger domain, RING-like zinc finger and SNF2 family N-terminal domain, B-box zinc finger, CBL proto-oncogene N-terminus EF hand-like domain, CBL proto-oncogene SH2-like domain and CBL proto-oncogene N-terminal domain 1 are shown dark pink, dark gray, pine green, denim blue, rose, orange, plum, light green, the darkest purple, blossom pink, lemon chiffon, yellow green, burgundy, blood red and red wine.

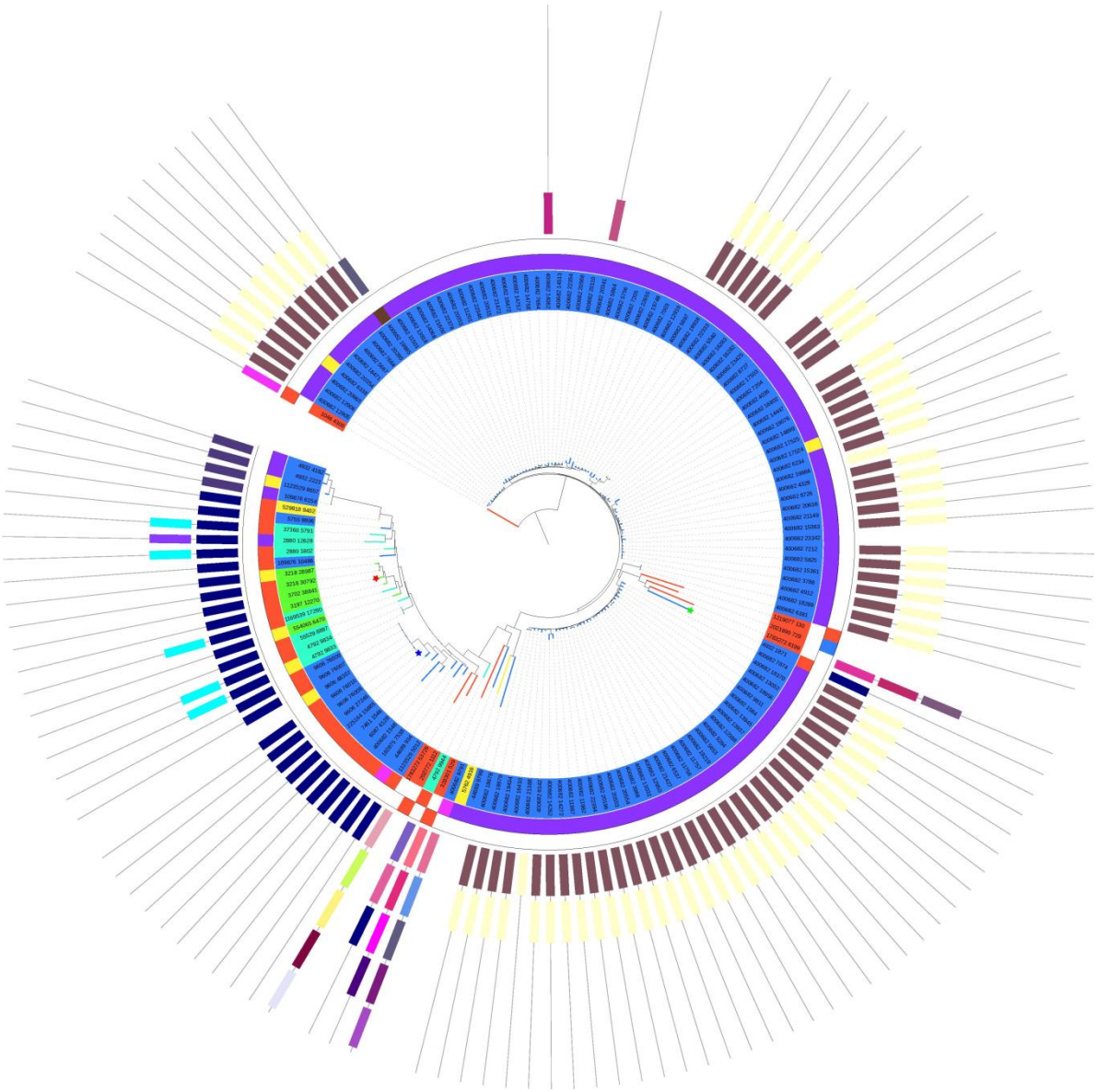


Figure 32. Phylogenetic tree of the protein Peroxisomal membrane protein (Pex14) reconstructed from the protein sequences that were detected in the HMM search with profiles which were built from the orthologous protein sequences of eukaryotic organisms that are part of the database. Initially used query sequences from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are denoted with blue, green and red star respectively. In total 28 domains were found and the most common are Peroxisomal membrane anchor protein (Pex14p) conserved region, PUB domain, protein tyrosine kinase domain and protein kinase domain shown in navy blue, cyan, cream and dull purple respectively.

4. Discussion

Proteins that make ancestral eukaryotic peroxisomal proteome can be divided into four groups according to their function. First group contains only catalase a hallmark peroxisomal enzyme. Second group comprises of Pxa1 and Pxa2 that are involved in lipid transport more precisely in the transport of very long chain acyl-CoA (Morita and Imanaka, 2012). Third group are proteins that are part of the peroxisomal import machinery which are Pex1, Pex2, Pex4, Pex5, Pex10 and Pex14, while fourth group includes Fox1p, Fox2p and Faa2p which are all involved in peroxisomal β-oxidation.

4.1. Catalase

Catalase breaks down H_2O_2 that is a metabolic by-product of many oxidase enzymes in order to maintain oxidative balance inside of peroxisome. In Figure 4. eukaryotic catalases are divided into three distinctive clades. First one contains catalases from superkingdom Unikonta, i.e. from phyla Amoebozoa, Filasteria and Fungi. While in all other catalase sequences mostly catalase and catalase-related immune-responsive domain was detected, proteins from this clade also contain domain from DJ-1/PfpI family which is usually associated with proteases and transcriptional regulators (Mitchell et al., 2018). Second clade consists of catalases from superkingdom Plantae and the third clade consists of catalase sequences from all four eukaryotic superkingdoms and Asgard group that is considered to be the closest prokaryotic relative of eukaryotes (Eme et al., 2017).

Second phylogenetic tree (Figure 21.) comprises of more prokaryotic sequences reveals one more clade with proteins from *Amphimedon queenslandica* (sponge) that contain AIG1 domain, 50S ribosome-binding GTPase domain and Interferon-inducible GTPase which all have role of GTP binding and are linked to self defense, interaction with 50S ribosome and intracellular defense respectively (Mitchell et al., 2018). Long branch length of this clade in comparison to the others and the absence of catalase domain indicates very distant relationship or even an error in HMM search. Other three clades are the same as ones detected in the first catalase tree but with more prokaryotic sequences grouped with them. Most common bacterial phyla in each clade are Firmicutes, Proteobacteria and Actinobacteria. Three distinctive clades of monofunctional catalase were revealed in previous studies (Klotz, Klassen and Loewen, 1997; Zamocky, Furtmüller and Obinger, 2008; Zámocký et al., 2012) and they are highly similar to this one with several smaller differences. First clade besides the sequences from Fungi and Bacteria contains sequences from eukaryotic phyla Amoebozoa and Filasteria. Member from Amoebozoa, *Dictyostelium discoideum*, was noticed by Zámocký et al., 2012 and it was explained as a result of

horizontal gene transfer due to the fact that is very frequent for Dictyostelium to receive genetic material from bacteria living in the same environment. Here, protein from an additional member of the same phylum is detected together with Dictyostellium and that is a protein from *Acanthamoeba castellanii* that has the same protein domain composition just as fungi *Gonapodya prolifera* that is considered to be a standard member of this clade. Nevertheless, catalase from *Capsaspora owczarzaki*, that is a member of phylum Filasteria, is part of this clade which is quite interesting considering the fact that is a one of the closest unicellular relatives to animals (Suga et al., 2013).

4.2. Lipid transport

Pxa1 and Pxa2 are part of the ATP-binding cassette (ABC) transporter subfamily D member 1 and member 2 respectively. Their peroxisomal phylogenetic trees (Figure 16. and Figure 17.) are nearly identical. Eukaryotic proteins can roughly be divided into four clades. First clade contains proteins from superkingdom Plantae and Chromalveolates that are grouped with bacterial proteins from Cyanobacteria and Proteobacteria. Domain from SbmA/BacA-like family distinguishes proteins of this clade from the rest of the tree. An *Escherichia coli* homologue of this domain is implicated in the uptake of microcins and bleomycin. This family is likely considered to be a subfamily of the ABC transporter family (Mitchell et al., 2018). Eukaryotic proteins of this clade are predicted to mostly reside in ER. Second clade contains proteins from superkingdoms Unikonta, Chromalveolates and Excavates that belong to the ABC transporter subfamily D member 4. Members of this clade are predicted to be localized in ER which at first does not make sense since ABCD4 transporters are localized to lysosome and take part in the transport of vitamin B12 from lysosome to the cytosol (Kawaguchi and Morita, 2016a). However, ABCD4 transporters are quite involved with endoplasmic reticulum before they end up as a part of the lysosome. After their translation on free polysomes, they are recognized by certain signal recognition particles and integrated into the ER membrane and then translocated to lysosome through an interaction with the lysosomal membrane protein LMBD1 (Kawaguchi et al., 2016b). Third clade contains proteins from ABC transporters subfamily D member 1 and member 2. Most members of this clade are predicted to be localized in mitochondria, although it is known that there are peroxisomal. Predictors mistake can be justified since those proteins do not contain typical peroxisomal targeting signals like PTS1 or PTS2, but mPTS (membrane peroxisomal targeting signal) has been identified in both groups, however no experimental data are available to support its functionality (Halbach et al., 2005). Fourth clade includes proteins from ABC transporter subfamily D member 3. ABCD3 is involved in the transport of branched chain acyl-CoA into peroxisomes. Even though mPTs have been identified

(Kashiwayama et al., 2007), members of this clade have also been predicted to reside in mitochondria.

Phylogenetic tree of Pxa2 (Figure 23.) is nearly the same as ones discussed previously, while orthology tree of Pxa1 shown in Figure 22. has two more eukaryotic clades. First clade contains eukaryotic proteins that are part of ABC transporter subfamily B member 6 and member 7 and proteobacterial sequences. Eukaryotic members of this clade are predicted to reside inside mitochondrion which is known for members of ABCB6 and ABCB7 (Krishnamurthy et al., 2006). Second clade contains multi-drug resistance proteins and members from ABCB4 which are predicted to the cell membrane and that agrees with the experimental data (Morita et al., 2013). All three eukaryotic clades (mitochondrial, cell membrane and peroxisomal) are clustered with bacterial homologous from several different phyla which indicates independent origin of each group of ABC transporters.

4.3. Peroxisomal protein import

From previous studies it is known that five out of six most ancient peroxins that are involved in peroxisomal protein import show homology with ERAD system (Gabaldón et al., 2006; Schlüter et al., 2006). Peroxisomal phylogenetic tree of peroxisomal biogenesis factor 1 (Pex1) (Figure 19.) has two eukaryotic clades. First one is Pex6 and the other one is Pex1. They both contain ATPase family associated with various cellular activities (AAA) domain, but Pex1 also has a Pex1 N-terminal domain. Members of both clades are predicted to cytoplasm which is accurate since Pex1 and Pex6 have dual localization during the release of the polyubiquitylated Pex5 receptor from the peroxisomal membrane in which they take action (Tamura et al., 2006).

Phylogenetic tree of Pex1 (Figure 27.) contains four more clades that do not appear in the peroxisomal tree. First clade represents nuclear valosin-containing protein-like that is predicted to reside within nucleus and it is involved in various nuclear processes such as assembly of the telomerase holoenzyme (Her and Chung, 2012), early and late stages of the pre-rRNA processing pathway (Yoshikatsu et al., 2015) etc. This protein belongs to the AAA ATPase family same as Pex1. Second clade contains Cell division control protein 48 (CDC48) and Transitional endoplasmic reticulum ATpase which have role in ERAD (Jarosch et al., 2002; Ye et al., 2004). Third clade contains spermatogenesis-associated protein 5 which also belongs to the AAA ATPase family and it is an ATP-dependent chaperone which uses the energy provided by ATP hydrolysis to generate mechanical force to disassemble protein complexes (Zakalskiy et al., 2002). Fourth clade is actinobacterial and at first it leads to the conclusion that both Pex1 and Pex6 have actinobacterial origin, but that was proven wrong. This was firstly noticed by Narendra et al., 2009 which proposed actinobacterial origin

of peroxisome since Pex1 and Pex6 are closer in distance to CDC48 homologs in Actinobacteria than to ER-localized CDC48 (Narendra et al., 2009). When this data was re-examine, it was proven that phylogenetic clustering of Pex1/Pex6 with actinobacterial homologs is most likely result of a long branch attraction artifact (Gabaldón and Capella-Gutiérrez, 2010). Based on this results Pex1 evolved from Cdc48 and Transitional endoplasmic reticulum ATPase which was previously proved by Gabaldón et al. but in the case of Pex2, Pex4, Pex5 and Pex10 the levels of sequence identity between the shared domains and the short regions of homology prevent the reconstruction of reliable phylogenetic trees to conclude that they also evolved from components of ERAD (Gabaldón et al., 2006).

Peroxisomal biogenesis factor 2 is a part of the ubiquitylation cascade that polyubiquitylates the Pex5 receptor (Smith and Aitchison, 2013). In its first phylogenetic tree (Figure 11.) two peroxisomal clades can be noticed. First one is Pex2 and the second one is Pex10. Members of those clades are predicted to reside in peroxisome and they have a Pex2/Pex12 amino terminal region which is characteristic for Pex2, Pex10 and Pex12 (Mitchell et al., 2018). Pex10 clade is grouped together with an ER clade whose members contain different Zn-RING finger domains such as zinc-RING finger domain, RING/Ubox like zinc-binding domain and RING-type zinc finger which are also present in two E3s ubiquitin ligases that participate in the ERAD process (Schlüter et al., 2006). This clade contains E3 ubiquitin-protein ligase RNF185 from several species for which it is known that it acts in ERAD (Kaneko et al., 2016) and E3 ubiquitin-protein ligase RNF5 that is involved in regulation of ERAD (Tcherpakov et al., 2009). Pex2 clade is clustered with two clades whose members are predicted to reside in nucleus and cytoplasm. Member of the first clade is human protein E3 ubiquitin-protein ligase RNF8 that plays a key role in DNA damage signaling (Ito et al., 2001). Member of the second clade is human protein E3 ubiquitin-protein ligase RNF146 that specifically binds poly-ADP-ribosylated (PARsylated) proteins and mediates their ubiquitination and subsequent degradation. It is involved in many biological processes, such as cell survival and DNA damage response (Zhang et al., 2011).
In the second phylogenetic tree of Pex2 (Figure 28.) same observations for Pex2 and Pex10 can be noticed.

Peroxin 4 is part of the ubiquitylation machinery in the alternative receptor cycling pathway of Pex5 receptor (Smith and Aitchison, 2013). In Figure 12. phylogenetic tree of Pex4 is shown and query Pex4 sequences from *Arabidopsis thaliana* and *Saccharomyces cerevisiae* are located in one clade while query from *Homo sapiens* in another. Arabidopsis/Saccharomyces clade contains almost all members from superkingdoms Plantae, Chromalveolates and Excavates while from Unikonta most of the representatives are missing. In the MetaPhors database there are listed orthologous sequences of five

species from Unikonta out of fifteen that are used in this analysis (Pryszcz, Huerta-Cepas and Gabaldón, 2010) and according to Schlüter et al., 2006 in some eukaryotic lineages their distribution is restricted. Pex4-conjugating enzyme family E2 shows homology with E2 components of ER. HMM profile search detected homology with various E2 components that are localized in nucleus and involved in different biological process which resulted in a large phylogenetic tree. ER homology was also detected but ER clade, that contains ubiquitin-conjugating enzyme E2 J1 and ubiquitin-conjugating enzyme E2 J2 which both take part in ERAD, is quite distance from Pex4 clade. However, second phylogenetic tree of Pex4 (Figure 29.), that includes five hundred proteins with the highest E-value from the HMM search with the profile that was built out of orthologous protein sequences shows, close relationship of ER proteins and Pex2. ER clade, that contains ubiquitin-conjugating enzyme E2 6 (UBC6) from *Saccharomyces cerevisiae*, that is part of ERAD and for which homology was found with Pex4 (Schlüter et al., 2006) is an ancestor clade to Pex4. Pex4 clade is clustered with a nuclear clade that contains proteins which are involved in different protein ubiquitination pathways but interestingly one of them is ubiquitin-conjugating enzyme E2 G2 for which it is proven that takes part in ERAD pathway (Sato et al., 2012).

Peroxisomal targeting signal 1 receptor (Pex5) binds to the C-terminal PTS1 and plays an essential role in the peroxisomal protein import. Peroxisomal phylogenetic tree (Figure 20.) shows PEX5 clade clustered with a clade containing cell division cycle protein 23 and cell division cycle protein 27 which are both components of anaphase promoting complex/cyclosome (APC/C) that is cell cycle-regulated E3 ubiquitin ligases that control progression through mitosis and the G1 phase of the cell cycle (Jin et al., 2008). Members of that clade are predicted to reside in cytoplasm and beside various TRP domains also have Anaphase promoting complex subunit 8 and Anaphase promoting complex subunit 3 characteristic for CDC23 and CDC27 respectively. Based on this result Pex5 differs from other ancient peroxins that show evolutionary relationship with components of ERAD but that can be even more corroborated. Firstly, TPR repeat in PEX5 is classified in a different class of TPR repeats than a TPR repeat in HRD3 (proposed homolog of Pex5 in the ERAD system of *Saccharomyces cerevisiae*) (Gabaldón et al., 2006). Interestingly, anaphase promoting complex/cyclosome uses TPR repeat protein together with a protein containing an E2 ubiquitin conjugating enzyme and a RING domain just like ERAD but ERAD uses all those domains additionally with an AAA+ ATPase. Furthermore, all other ancient peroxins are directly associated to the peroxisomal membrane while Pex5 spends most of its time in the cytoplasm so it would make sense for peroxins that are in the peroxisomal membrane to arise from ER and it was experimentally confirmed that peroxisomal membrane emerges from ER membrane (Tabak et al., 2003) while Pex5 may emerge from the molecular system

localized in cytoplasm. This result indicates that peroxisomal import system originates not from one but two molecular systems, ERAD and Anaphase promoting complex.

Pex5 clade misses the query sequence from *Saccharomyces cerevisiae* that was clustered in a different clade of the eukaryotic phylogenetic tree (Figure 13.). That was probably caused because of the extremely high number of protein sequences (3688) used in the reconstruction of that tree and not by different origin. In the orthologous tree (Figure 30.) Pex5 from *Saccharomyces cerevisiae* made the cut in the homology search and it is part of the Pex5 clade which discards the possibility of an independent origin.

Peroxisome biogenesis factor 10 (Pex10) together with Pex2 and Pex12 makes a polyubiquitylation cascade that polyubiquitylates the Pex5 receptor which is then released from the peroxisomal membrane (Smith and Aitchison, 2013). Peroxisomal clade can be easily detected in the first phylogenetic tree (Figure 14.) because all three query sequences are grouped together, all members are predicted to reside in peroxisome and proteins have Pex2/Pex12 amino terminal region which is characteristic for Pex10 as mentioned earlier. Other detected protein families in the tree belong to the E3 ubiquitin-protein ligase family which are localized in ER, nucleus and cytoplasm. Bigger clade that contains peroxisomal clade is grouped together with an ER clade which contains E3 ubiquitin-protein ligase RNF185 and E3 ubiquitin-protein ligase RNF5 that are involved in ERAD as mentioned and observed in Pex2 phylogenetic trees (Figure 11. and Figure 28.). In the second Pex10 tree (Figure 31.) peroxisomal clade can also be easily noticed but here two separated ER clades appear. Peroxisomal clade is clustered together with two clades. First clade contains helicase-like transcription factor which has both helicase and E3 ubiquitin ligase activity (Unk et al., 2008) and it mostly predicted to reside in nucleus. Second clade contains E3 ubiquitin-protein ligase RNF170 which is involved in ERAD (Lu et al., 2011) and cytoplasmic RING protein 32 that may play role in sperm formation (van Baren et al., 2002). This three clades are grouped with an ER clade that contains RING finger protein 145 which is involved in maintenance of cholesterol homeostasis (Zhang et al., 2017), E3 ubiquitin protein ligase synoviolin a precursor and E3 ubiquitin protein ligase AMFR that both participate in ERAD (Nadav et al., 2003; Fang et al., 2001). Even though different homology search method was used in each Pex10 tree both of them indicate strong evolutionary relationship with components of ERAD.

Peroxisomal membrane protein Pex14 has a role as a docking factor for the Pex5 receptor (Fransen, Terlecky and Subramani, 1998). Out of all analyzed proteins Pex14 by far had the least number of hits in the HMM profile search (Table 2. and Table 11.). In the first phylogenetic tree (Figure 15.) there are no other protein families besides the Pex14 one. It is an only ancient peroxin that does not show homology with components of ERAD (Gabaldón et al., 2006). Pex14 are roughly clustered in two main clades where first one contains

members from Unikonta while the second one has members from other three eukaryotic superkingdoms and also members from phylum Amoebozoa which is considered to be a sister group to animals and fungi (Eichinger et al., 2005). Based on this phylogenetic clustering there was probably a duplication of Pex14 in LECA that caused this dichotomy. Second phylogenetic tree of Pex14 (Figure 32.) surprisingly has 103 protein sequences with protein tyrosine kinase domain and/or protein kinase domain from *Amphimedon queenslandica* (sponge). In MetaPhors database from which orthologous sequences were retrieved for building an HMM profile, two orthologs of Pex14 from sponge are present. First one is peroxisomal membrane protein and the second one is tip elongation aberrant protein 3-like that contains both domains that were detected in 103 protein sequences mentioned previously. There were a lot of isoforms of the same protein detected which is not surprising since the proteome of *Amphimedon queenslandica* is quite large and has 43,435 proteins (Srivastava et al., 2010). Besides that anomaly the same dichotomy in Pex14 clade can be noticed.

4.4. Peroxisomal β-oxidation

Last group of proteins are ones associated with peroxisomal β-oxidation. Long-chain fatty acid ligase 2 (Faa2) converts long-chain fatty acids into metabolically active CoA thioesters that can either be degraded via peroxisomal beta-oxidation or incorporated into phospholipids. Peroxisomal tree of Faa2 (Figure 19.) has three eukaryotic clades. First eukaryotic clade has long-chain-fatty-acid--CoA ligase ACSBG1 and long-chain-fatty-acid--CoA ligase ACSBG2, which belong to the so called “bubblegum” family capable of activating very long-chain fatty acids. Most members of this clade are predicted to reside in peroxisome and in human counterpart PTS2 (RIDPSCPQL) was found at amino acid residue 93. But by using indirect immunofluorescence and confocal microscopy it was shown that human protein appeared to be close to or associated with the plasma membrane (Steinberg et al., 2000). Second eukaryotic clade that is mostly peroxisomal contains long-chain-fatty-acid--CoA ligase 4 (ACSL4), for whom it was proved to be highly expressed in liver peroxisomes but none of the mammalian ACSL isoforms contain PTS1 or PTS2 (Watkins and Ellis, 2012), which probably explains why some members of this clade were predicted to reside in ER. Third eukaryotic clade contains query proteins from *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*, surprisingly a lot of members of this clade are predicted to reside in ER. ACSL5 and ACSL1 are located in this clade. ACSL5 was found in ER and ACSL 1 in peroxisome by proteomics (Islinger et al., 2010). Orthologous phylogenetic tree of Faa2 (Figure 26.) is the same as the peroxisomal but only without the first eukaryotic clade.

Interestingly, both trees for which different homology search approach was used don't show any evolutionary relationship with mitochondrial counterparts.

Second protein from this group is acyl-coenzyme A oxidase (Fox1) that catalyses the first step of the peroxisomal β-oxidation which catalyzes the desaturation of acyl-CoAs to 2-trans-enoyl-CoAs and donates electrons directly to molecular oxygen, thereby producing hydrogen peroxide. In the first phylogenetic tree of Fox1 (Figure 7.) it is quite easy to notice the split between the peroxisomal and mitochondrial counterparts. Midpoint rooting method was applied instead of using an outgroup because different choice of bacterial clade as an outgroup resulted in different tree topologies and by using midpoint rooting any possible bias was avoided. Peroxisomal clade contains acyl-coenzyme A oxidase-like protein, acyl-coenzyme A oxidase 1, acyl-coenzyme A oxidase 2 and acyl-coenzyme A oxidase 3. Mitochondrial clade contains acyl-CoA dehydrogenase family member 10, acyl-CoA dehydrogenase family member 11, glutaryl-Coa dehydrogenase, isovaleryl-CoA dehydrogenase, very long-chain specific acyl-CoA dehydrogenase, short/branched chain specific acyl-CoA dehydrogenase, short-chain specific acyl-CoA dehydrogenase, medium-chain specific acyl-CoA dehydrogenase and isobutyryl-CoA dehydrogenase. In the mitochondrial part of the tree the members of the first clade that contains acyl-CoA dehydrogenase family member 10 and 11 are predicted to reside in peroxisome. These proteins do not have mitochondrial targeting sequences at their N-termini and therefore are likely to be localized to alternative cellular locations. In mouse ACAD10 protein is predicted to the membrane and to mitochondria (Kislinger et al., 2006). ACAD11 protein has a peroxisomal targeting signal at its C terminus (Kikuchi et al., 2003). Some findings suggest that some isomers may be localized to membrane associated vesicles, while in human neuroblastoma cells it is localized in mitochondria (He et al., 2011). These results indicate distinct and independent origin of peroxisomal acyl-CoA oxidase and mitochondrial acyl-CoA dehydrogenase. Fox1 was a main protein which supported Speijier's model which explains the origin of peroxisome. According to that model Fox1p has an ancestral mitochondrial location from where it was firstly retargeted to endomembrane system and afterwards it separated into new organelle, i.e. peroxisome (Speijer, 2013). Other proposed model by Gabaldón doesn't contain the first step and instead Fox1 was initially part of the endomembrane system from which peroxisome arisen (Gabaldón, 2014) and obtained phylogenetic tree of Fox1 agrees with this model. This model is also supported by the fact that Fox1 is involved in some desaturation steps in the synthesis of polyunsaturated fatty acids, a process partially located in ER (Gabaldón, Ginger and Michels, 2016). Interestingly, acyl-CoA dehydrogenase from Asgard group only appears in mitochondrial part of the tree, i.e. in 4 mitochondrial clades which are acyl-CoA dehydrogenase family member 10 and 11,

isovaleryl-CoA dehydrogenase, short-chain specific acyl-CoA dehydrogenase and medium-chain specific acyl-CoA dehydrogenase .

Peroxisomal hydratase-dehydrogenase-epimerase (Fox2) acts in the second step of beta-oxidation and it converts trans-2-enoyl-CoA via D-3-hydroxyacyl-CoA to 3-ketoacyl-CoA. First phylogenetic tree of Fox2 (Figure 8.) has one eukaryotic clade that represents Fox2 and it is clustered with proteins from Asgard group. Other eukaryotic proteins that are not within this clade are all mostly involved in fatty acid synthesis and they don't form a clade. They were detected in the HMM search because Fox2 contains MaoC-like domain which shares similarity with variety of enzymes among which is fatty acid synthase beta subunit (Mitchell et al., 2018). Interestingly, mitochondrial counterparts of this enzyme were not detected in the HMM profile search. In the second HMM search where profiles were built from orthologous protein sequences most of the hits with the highest e-value were bacterial so query sequence from *Arabidopsis thaliana* didn't even make the cut to be in the phylogenetic tree. High amount and high ranking of bacterial homologs can be justified since Fox2 has alphaproteobacterial origin (Gabaldón et al, 2006) and contains Maoc-like domain which is present in many bacteria. Furthermore, its alphaproetobacterial origin was again confirmed and origin from mitochondrial genome was proposed (Bolte, Rensing and Maier, 2014).

5. Conclusion

After the detailed phylogenetic analysis of the ancestral eukaryotic peroxisomal proteome was conducted, results were visualized and extensive discussion was done following conclusions can be made:

1. Adding members from the recently sequenced eukaryotic phyla and recently discovered Asgard group does not change the previously known tree topology of monofunctional catalase.
2. Pxa1 and Pxa2, that are involved in long-chain fatty acid transport show common origin with ATP-binding cassette subfamily D members (ABCD4) that are prior their final localization in lysosome part of the ER. This may indicate that Pxa1 and Pxa2 are also part of the ER before they end up in the peroxisomal membrane. Their mitochondrial and cell membrane counterparts were shown to have independent origin, which excludes their evolutionary connection with them.
3. Ancient peroxins Pex1, Pex2, Pe4 and Pex10 evolved from the homologous components of ERAD. Pex5 evolved from the cell division cycle protein 23 and cell division cycle protein 27 that are part of the anaphase promoting complex/cyclosome (APC/C) while Pex14 doesn't show evolutionary relationship to those two molecular system or any other and it is probably a novel protein. According to this results peroxisomal protein import machinery originated from two molecular systems: ERAD and anaphase promoting complex/cyclosome (APC/C).
4. Fox1, that catalyzes the first step of peroxisomal beta oxidation, has an independent origin from his mitochondrial counterpart which confirms Gabaldón's model on the origin of peroxisome. For Fox2 it was not possible to conduct a proper homology search since there was too many bacterial hits with very high e-values in both HMM profile searches which are probably caused by its alphaproteobacterial origin that was previously proven. Faa2 doesn't have common origin with its mitochondrial counterparts and shows strong relationship with ER since most of the detected homologs across different eukaryotic phyla reside either in ER or peroxisome.

6. References

Almagro Armenteros, J., Sønderby, C., Sønderby, S., Nielsen, H. and Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(24), pp.4049-4049.

Bolte, K., Rensing, S. and Maier, U. (2014). The evolution of eukaryotic cells from the perspective of peroxisomes. *BioEssays*, 37(2), pp.195-203.

Breidenbach, R., Kahn, A. and Beevers, H. (1968). Characterization of Glyoxysomes From Castor Bean Endosperm. *PLANT PHYSIOLOGY*, 43(5), pp.705-713.

Capella-Gutierrez, S., Silla-Martinez, J. and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15), pp.1972-1973.
Cavalier-Smith, T. (1987). The simultaneous symbiotic origin of mitochondria, chloroplasts and microbodies. Ann. N.Y. Acad. Sci. 503,55–71.

Cavalier-Smith, T. (1997). Cell and genome co-evolution: facultative anaerobiosis, glycosomes and kinetoplastan RNA editing. Trends Genet. 13, 6–9.

Cooper, G. M., & Hausman, R. E. (2009). *The cell: a molecular approach*. Washington, D.C., ASM Press.

De Duve, C. and Baudhuin, P. (1966). Peroxisomes (microbodies and related particles). *Physiological Reviews*, 46(2), pp.323-357.

de Duve C. Evolution of the peroxisome. Ann NY Acad Sci 1969;168:369-81

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis.*

Eddy, S. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, 6(3), pp.361-365.

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9), pp.755-763.

Eddy, S. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10), p.e1002195.

Eddy, S. (2018). *Profile HMM Analysis*. [online] Biology.wustl.edu. Available at: http://www.biology.wustl.edu/gcg/hmmanalysis.html [Accessed 30 Aug. 2018].

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), pp.1792-1797.

Eichinger, L., Pachebat, J., Glöckner, G., Rajandream, M., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., Tunggal, B., Kummerfeld, S., Madera, M., Konfortov, B., Rivero, F., Bankier, A., Lehmann, R., Hamlin, N., Davies, R., Gaudet, P., Fey, P., Pilcher, K., Chen, G., Saunders, D., Sodergren, E., Davis, P., Kerhornou, A., Nie, X., Hall, N., Anjard, C., Hemphill, L., Bason, N., Farbrother, P., Desany, B., Just, E., Morio, T., Rost, R., Churcher, C., Cooper, J., Haydock, S., van Driessche, N., Cronin, A., Goodhead, I., Muzny, D., Mourier, T., Pain, A., Lu, M., Harper, D., Lindsay, R., Hauser, H., James, K., Quiles, M., Madan Babu, M., Saito, T., Buchrieser, C., Wardroper, A., Felder, M., Thangavelu, M., Johnson, D., Knights, A., Loulseged, H., Mungall, K., Oliver, K., Price, C., Quail, M., Urushihara, H., Hernandez, J., Rabbinowitsch, E., Steffen, D., Sanders, M., Ma, J., Kohara, Y., Sharp, S., Simmonds, M., Spiegler, S., Tivey, A., Sugano, S., White, B., Walker, D., Woodward, J., Winckler, T., Tanaka, Y., Shaulsky, G., Schleicher, M., Weinstock, G., Rosenthal, A., Cox, E., Chisholm, R., Gibbs, R., Loomis, W., Platzer, M., Kay, R., Williams, J., Dear, P., Noegel, A., Barrell, B. and Kuspa, A. (2005). The genome of the social amoeba Dictyostelium discoideum. *Nature*, 435(7038), pp.43-57.

Eme, L., Spang, A., Lombard, J., Stairs, C. and Ettema, T. (2017). Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*, 15(12), pp.711-723.

Enright, A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7), pp.1575-1584.

Erdmann, R. and Kunau, W. (1992). A genetic approach to the biogenesis of peroxisomes in the yeast Saccharomyces cerevisiae. *Cell Biochemistry and Function*, 10(3), pp.167-174.

Fang, S., Ferrone, M., Yang, C., Jensen, J., Tiwari, S. and Weissman, A. (2001). The tumor autocrine motility factor receptor, gp78, is a ubiquitin protein ligase implicated in degradation from the endoplasmic reticulum. *Proceedings of the National Academy of Sciences*, 98(25), pp.14422-14427.

Fransen, M., Terlecky, S. and Subramani, S. (1998). Identification of a human PTS1 receptor docking protein directly required for peroxisomal protein import. *Proceedings of the National Academy of Sciences*, 95(14), pp.8087-8092.

Freitag, J., Ast, J. and Bölker, M. (2012). Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. *Nature*, 485(7399), pp.522-525.

Gabaldón, T., Snel, B., Zimmeren, F., Hemrika, W., Tabak, H. and Huynen, M. (2006). *Biology Direct*, 1(1), p.8.

Gabaldón, T. (2010). Peroxisome diversity and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1541), pp.765-773.

Gabaldón, T. (2014). Evolutionary considerations on the origin of peroxisomes from the endoplasmic reticulum, and their relationships with mitochondria. *Cellular and Molecular Life Sciences*, 71(13), pp.2379-2382.

Gabaldón, T. and Pittis, A. (2015). Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes. *Biochimie*, 119, pp.262-268.

Gabaldón, T., Ginger, M. and Michels, P. (2016). Peroxisomes in parasitic protists. *Molecular and Biochemical Parasitology*, 209(1-2), pp.35-45.

Hajra, A. (1995). Glycerolipid biosynthesis in peroxisomes (microbodies). *Progress in Lipid Research*, 34(4), pp.343-364.

Halbach, A., Lorenzen, S., Landgraf, C., Volkmer-Engert, R., Erdmann, R. and Rottensteiner, H. (2005). Function of the PEX19-binding Site of Human Adrenoleukodystrophy Protein as Targeting Motif in Man and Yeast. *Journal of Biological Chemistry*, 280(22), pp.21176-21182.

He, M., Pei, Z., Mohsen, A., Watkins, P., Murdoch, G., Van Veldhoven, P., Ensenauer, R. and Vockley, J. (2011). Identification and characterization of new long chain Acyl-CoA dehydrogenases. *Molecular Genetics and Metabolism*, 102(4), pp.418-429.

Her, J. and Chung, I. (2012). The AAA-ATPase NVL2 is a telomerase component essential for holoenzyme assembly. *Biochemical and Biophysical Research Communications*, 417(3), pp.1086-1092.

Hoepfner, D., Schildknegt, D., Braakman, I., Philippsen, P. and Tabak, H. (2005). Contribution of the Endoplasmic Reticulum to Peroxisome Formation. *Cell*, 122(1), pp.85-95.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L., Denisov, I., Kormes, D., Marcet-Houben, M. and Gabaldon, T. (2010). PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Research*, 39(Database), pp.D556-D560.

Islinger, M., Li, K., Loos, M., Liebler, S., Angermüller, S., Eckerskorn, C., Weber, G., Abdolzade, A. and Völkl, A. (2010). Peroxisomes from the Heavy Mitochondrial Fraction: Isolation by Zonal Free Flow Electrophoresis and Quantitative Mass Spectrometrical Characterization. *Journal of Proteome Research*, 9(1), pp.113-124.

Ito, K., Adachi, S., Iwakami, R., Yasuda, H., Muto, Y., Seki, N. and Okano, Y. (2001). N-Terminally extended human ubiquitin-conjugating enzymes (E2s) mediate the ubiquitination of RING-finger proteins, ARA54 and RNF8. *European Journal of Biochemistry*, 268(9), pp.2725-2732.

Jarosch, E., Taxis, C., Volkwein, C., Bordallo, J., Finley, D., Wolf, D. and Sommer, T. (2002). Protein dislocation from the ER requires polyubiquitination and the AAA-ATPase Cdc48. *Nature Cell Biology*, 4(2), pp.134-139.

Jedd, G. and Chua, N. (2010). *A new self-assembled peroxisomal vesicle required for efficient resealing of the plasma membrane*. Nat Cell Biol, 2, 226-31

Jin, L., Williamson, A., Banerjee, S., Philipp, I. and Rape, M. (2008). Mechanism of Ubiquitin-Chain Formation by the Human Anaphase-Promoting Complex. *Cell*, 133(4), pp.653-665.

Jung, S., Smith, J., von Haller, P., Dilworth, D., Sitko, K., Miller, L., Saleem, R., Goodlett, D. and Aitchison, J. (2013). Global Analysis of Condition-specific Subcellular Protein Distribution and Abundance. *Molecular & Cellular Proteomics*, 12(5), pp.1421-1435.

Kaneko, M., Iwase, I., Yamasaki, Y., Takai, T., Wu, Y., Kanemoto, S., Matsuhisa, K., Asada, R., Okuma, Y., Watanabe, T., Imaizumi, K. and Nomura, Y. (2016). Genome-wide identification and gene expression profiling of ubiquitin ligases for endoplasmic reticulum protein degradation. *Scientific Reports*, 6(1).

Kashiwayama, Y., Asahina, K., Morita, M. and Imanaka, T. (2007). Hydrophobic Regions Adjacent to Transmembrane Domains 1 and 5 Are Important for the Targeting of the 70-kDa Peroxisomal Membrane Protein. *Journal of Biological Chemistry*, 282(46), pp.33831-33844.

Katoh,K., Asimenos,G. and Toh,H. (2009) Multiple alignment of DNA sequences with MAFFT. Methods Mol. Biol., 537 , 39–64.
Kawaguchi, K. and Morita, M. (2016). ABC Transporter Subfamily D: Distinct Differences in Behavior between ABCD1–3 and ABCD4 in Subcellular Localization, Function, and Human Disease. *BioMed Research International*, 2016, pp.1-11.

Kawaguchi, K., Okamoto, T., Morita, M. and Imanaka, T. (2016). Translocation of the ABC transporter ABCD4 from the endoplasmic reticulum to lysosomes requires the escort protein LMBD1. *Scientific Reports*, 6(1).

Kikuchi, M., Hatano, N., Yokota, S., Shimozawa, N., Imanaka, T. and Taniguchi, H. (2003). Proteomic Analysis of Rat Liver Peroxisome. *Journal of Biological Chemistry*, 279(1), pp.421-428.

Kislinger, T., Cox, B., Kannan, A., Chung, C., Hu, P., Ignatchenko, A., Scott, M., Gramolini, A., Morris, Q., Hallett, M., Rossant, J., Hughes, T., Frey, B. and Emili, A. (2006). Global Survey of Organ and Organelle Protein Expression in Mouse: Combined Proteomic and Transcriptomic Profiling. *Cell*, 125(1), pp.173-186.

Klotz, M., Klassen, G. and Loewen, P. (1997). Phylogenetic relationships among prokaryotic and eukaryotic catalases. *Molecular Biology and Evolution*, 14(9), pp.951-958.

Krishnamurthy, P., Du, G., Fukuda, Y., Sun, D., Sampath, J., Mercer, K., Wang, J., Sosa-Pineda, B., Murti, K. and Schuetz, J. (2006). Identification of a mammalian mitochondrial porphyrin transporter. *Nature*.

Kruger NJ, von Schaewen A. The oxidative pentose phosphate pathway: structure and organisation. Current Opinion in Plant Biology. 2003;6:236–246

Lassmann, T., Frings, O. and Sonnhammer, E. (2008). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Research*, 37(3), pp.858-865.

Lazarow, P. and Fujiki, Y. (1985). Biogenesis of Peroxisomes. *Annual Review of Cell Biology*, 1(1), pp.489-530.

Letunic, I. and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, 44(W1), pp.W242-W245.

Lodhi, I. and Semenkovich, C. (2014). Peroxisomes: A Nexus for Lipid Metabolism and Cellular Signaling. *Cell Metabolism*, 19(3), pp.380-392.

Losón, O., Song, Z., Chen, H. and Chan, D. (2013). Fis1, Mff, MiD49, and MiD51 mediate Drp1 recruitment in mitochondrial fission. *Molecular Biology of the Cell*, 24(5), pp.659-667.

Lu, J., Wang, Y., Sliter, D., Pearce, M. and Wojcikiewicz, R. (2011). RNF170 Protein, an Endoplasmic Reticulum Membrane Ubiquitin Ligase, Mediates Inositol 1,4,5-Trisphosphate Receptor Ubiquitination and Degradation. *Journal of Biological Chemistry*, 286(27), pp.24426-24433.

Meinken, J. and Min, J. (2012). Computational Prediction of Protein Subcellular Locations in Eukaryotes: an Experience Report. *Computational Molecular Biology*.

Mitchell, A., Attwood, T., Babbitt, P., Blum, M., Bork, P., Bridge, A., Brown, S., Chang, H., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D., Necci, M., Nuka, G., Orengo, C., Pandurangan, A., Paysan-Lafosse, T., Pesseat, S., Potter, S., Qureshi, M., Rawlings, N., Redaschi, N., Richardson, L., Rivoire, C., Salazar, G., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Sutton, G., Thanki, N., Thomas, P., Tosatto, S., Yong, S. and Finn, R. (2018). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*.

Morita, M. and Imanaka, T. (2012). Peroxisomal ABC transporters: Structure, function and role in disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1822(9), pp.1387-1396.

Morita, S., Tsuda, T., Horikami, M., Teraoka, R., Kitagawa, S. and Terada, T. (2013). Bile salt-stimulated phospholipid efflux mediated by ABCB4 localized in nonraft membranes. *Journal of Lipid Research*, 54(5), pp.1221-1230.

Motley, A. and Hettema, E. (2007). Yeast peroxisomes multiply by growth and division. *The Journal of Cell Biology*, 178(3), pp.399-410.

Motley, A., Ward, G. and Hettema, E. (2008). Dnm1p-dependent peroxisome fission requires Caf4p, Mdv1p and Fis1p. *Journal of Cell Science*, 121(10), pp.1633-1640.

Mount, D. (2006). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Muller, M., Mentel, M., van Hellemond, J., Henze, K., Woehle, C., Gould, S., Yu, R., van der Giezen, M., Tielens, A. and Martin, W. (2012). Biochemistry and Evolution of Anaerobic Energy Metabolism in Eukaryotes. *Microbiology and Molecular Biology Reviews*, 76(2), pp.444-495.

Nadav, E., Shmueli, A., Barr, H., Gonen, H., Ciechanover, A. and Reiss, Y. (2003). A novel mammalian endoplasmic reticulum ubiquitin ligase homologous to the yeast Hrd1. *Biochemical and Biophysical Research Communications*, 303(1), pp.91-97.

Narendra, D., Satoshi, S., Kazuo, H., Daisuke, M., Tokumasa, H. and Takao, S. (2009). A Study on the Origin of Peroxisomes: Possibility of Actinobacteria Symbiosis. *Nature Precedings*.

Novikoff, A. & Shin, W. Y., (1964) The endoplasmic reticulum in the Golgi zone and its relation to microbodies, Golgi apparatus and autophagic vacuoles in rat liver cells. J.Microsc. 3, 187–206.

Opaliński, Ł., Kiel, J., Williams, C., Veenhuis, M. and van der Klei, I. (2010). Membrane curvature during peroxisome fission requires Pex11. *The EMBO Journal*, 30(1), pp.5-16. *Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK.

Price, M., Dehal, P. and Arkin, A. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3), p.e9490.

Pryszcz, L., Huerta-Cepas, J. and Gabaldón, T. (2010). MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Research*, 39(5), pp.e32-e32.

Punta, M., Coggill, P., Eberhardt, R., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E., Eddy, S., Bateman, A. and Finn, R. (2011). The Pfam protein families database. *Nucleic Acids Research*, 40(D1), pp.D290-D301.

Rhodin, J., (1954). Correlation of ultrastructural organization and function in normal and experimentally changed proximal convoluted tubule cells of the mouse kidney. PhD dissertation, Aktiebolaget Godvil, Stockholm.

Rybicka, K. (1996). Glycosomes — the organelles of glycogen metabolism. *Tissue and Cell*, 28(3), pp.253-265.

Sato, T., Sako, Y., Sho, M., Momohara, M., Suico, M., Shuto, T., Nishitoh, H., Okiyoneda, T., Kokame, K., Kaneko, M., Taura, M., Miyata, M., Chosa, K., Koga, T., Morino-Koga, S., Wada, I. and Kai, H. (2012). STT3B-Dependent Posttranslational N-Glycosylation as a Surveillance System for Secretory Protein. *Molecular Cell*, 47(1), pp.99-110.

Schliebs, W., Girzalsky, W. and Erdmann, R. (2010). Peroxisomal protein import and ERAD: variations on a common theme. *Nature Reviews Molecular Cell Biology*, 11(12), pp.885-890.

Schlüter, A., Fourcade, S., Ripp, R., Mandel, J., Poch, O. and Pujol, A. (2006). The Evolutionary Origin of Peroxisomes: An ER-Peroxisome Connection. *Molecular Biology and Evolution*, 23(4), pp.838-845.

Schrader, M. and Fahimi, H. (2006). Peroxisomes and oxidative stress. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1763(12), pp.1755-1766.

Schrader, M., Bonekamp, N. and Islinger, M. (2012). Fission and proliferation of peroxisomes. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1822(9), pp.1343-1357.

Smith, J. and Aitchison, J. (2013). Peroxisomes take shape. *Nature Reviews Molecular Cell Biology*, 14(12), pp.803-817.

Speijer, D. (2013). Reconsidering ideas regarding the evolution of peroxisomes: the case for a mitochondrial connection. *Cellular and Molecular Life Sciences*, 71(13), pp.2377-2378.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M., Mitros, T., Richards, G., Conaco, C., Dacre, M., Hellsten, U., Larroux, C., Putnam, N., Stanke, M., Adamska, M., Darling, A., Degnan, S., Oakley, T., Plachetzki, D., Zhai, Y., Adamski, M., Calcino, A., Cummins, S., Goodstein, D., Harris, C., Jackson, D., Leys, S.,

Shu, S., Woodcroft, B., Vervoort, M., Kosik, K., Manning, G., Degnan, B. and Rokhsar, D. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*, 466(7307), pp.720-726.

Stabenau, H., Winkler, U., Säftel, W., 1998. On the origin of microbodies in plants. University of Oldenburg, Oldenburg, Germany, pp. 3-8.

Steinberg, S., Morgenthaler, J., Heinzer, A., Smith, K. and Watkins, P. (2000). Very Long-chain Acyl-CoA Synthetases. *Journal of Biological Chemistry*, 275(45), pp.35162-35169.

Suga, H., Chen, Z., de Mendoza, A., Sebé-Pedrós, A., Brown, M., Kramer, E., Carr, M., Kerner, P., Vervoort, M., Sánchez-Pons, N., Torruella, G., Derelle, R., Manning, G., Lang, B., Russ, C., Haas, B., Roger, A., Nusbaum, C. and Ruiz-Trillo, I. (2013). The Capsaspora genome reveals a complex unicellular prehistory of animals. *Nature Communications*, 4(1).

Sugiura, A., Mattie, S., Prudent, J. and McBride, H. (2017). Newly born peroxisomes are a hybrid of mitochondrial and ER-derived pre-peroxisomes. *Nature*, 542(7640), pp.251-254.

Tabak, H., Murk, J., Braakman, I. and Geuze, H. (2003). Peroxisomes Start Their Life in the Endoplasmic Reticulum. *Traffic*, 4(8), pp.512-518.

Tamura, S., Yasutake, S., Matsumoto, N. and Fujiki, Y. (2006). Dynamic and Functional Assembly of the AAA Peroxins, Pex1p and Pex6p, and Their Membrane Receptor Pex26p. *Journal of Biological Chemistry*, 281(38), pp.27693-27704.

Tcherpakov, M., Delaunay, A., Toth, J., Kadoya, T., Petroski, M. and Ronai, Z. (2009). Regulation of Endoplasmic Reticulum-associated Degradation by RNF5-dependent Ubiquitination of JNK-associated Membrane Protein (JAMP). *Journal of Biological Chemistry*, 284(18), pp.12099-12109.

Unk, I., Hajdu, I., Fatyol, K., Hurwitz, J., Yoon, J., Prakash, L., Prakash, S. and Haracska, L. (2008). Human HLTF functions as a ubiquitin ligase for proliferating cell nuclear antigen polyubiquitination. *Proceedings of the National Academy of Sciences*, 105(10), pp.3768-3773.

van Baren, M., van der Linde, H., Breedveld, G., Baarends, W., Rizzu, P., de Graaff, E., Oostra, B. and Heutink, P. (2002). A Double RING-H2 Domain in RNF32, a Gene Expressed during Sperm Formation. *Biochemical and Biophysical Research Communications*, 292(1), pp.58-65.

van der Klei, I. and Veenhuis, M. (2013). The Versatility of Peroxisome Function in Filamentous Fungi. *Peroxisomes and their Key Role in Cellular Signaling and Metabolism*, pp.135-152.

van der Zand, A., Gent, J., Braakman, I. and Tabak, H. (2012). Biochemically Distinct Vesicles from the Endoplasmic Reticulum Fuse to Form Peroxisomes. *Cell*, 149(2), pp.397-409.

van Dongen, S. and Abreu-Goodger, C. (2011). Using MCL to Extract Clusters from Networks. *Bacterial Molecular Networks*, pp.281-295.

Wallace, I., O'Sullivan, O., Higgins, D. and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, 34(6), pp.1692-1699.

Ye, Y., Shibata, Y., Yun, C., Ron, D. and Rapoport, T. (2004). A membrane protein complex mediates retro-translocation from the ER lumen into the cytosol. *Nature*, 429(6994), pp.841-847.

Yoshikatsu, Y., Ishida, Y., Sudo, H., Yuasa, K., Tsuji, A. and Nagahama, M. (2015). NVL2, a nucleolar AAA-ATPase, is associated with the nuclear exosome and is involved in pre-rRNA processing. *Biochemical and Biophysical Research Communications*, 464(3), pp.780-786.

Zakalskiy, A., Högenauer, G., Ishikawa, T., Wehrschütz-Sigl, E., Wendler, F., Teis, D., Zisser, G., Steven, A. and Bergler, H. (2002). Structural and Enzymatic Properties of the AAA Protein Drg1p from Saccharomyces cerevisiae. *Journal of Biological Chemistry*, 277(30), pp.26788-26795.

Zamocky, M., Furtmüller, P. and Obinger, C. (2008). Evolution of Catalases from Bacteria to Humans. *Antioxidants & Redox Signaling*, 10(9), pp.1527-1548.

Zámocký, M., Gasselhuber, B., Furtmüller, P. and Obinger, C. (2012). Molecular evolution of hydrogen peroxide degrading enzymes. *Archives of Biochemistry and Biophysics*, 525(2), pp.131-144.

Zhang, L., Rajbhandari, P., Priest, C., Sandhu, J., Wu, X., Temel, R., Castrillo, A., de Aguiar Vallim, T., Sallam, T. and Tontonoz, P. (2017). Inhibition of cholesterol biosynthesis through RNF145-dependent ubiquitination of SCAP. *eLife*, 6.

Zhang, Y., Liu, S., Mickanin, C., Feng, Y., Charlat, O., Michaud, G., Schirle, M., Shi, X., Hild, M., Bauer, A., Myer, V., Finan, P., Porter, J., Huang, S. and Cong, F. (2011). RNF146 is a poly(ADP-ribose)-directed E3 ligase that regulates axin degradation and Wnt signalling. *Nature Cell Biology*, 13(5), pp.623-629.

NCBI Genome Database: https://www.ncbi.nlm.nih.gov/genome

UniProt Proteomes Database: https://www.uniprot.org/proteomes/

7. Supplementary

Supplementary 1. List of eukaryotic species used in the database with corresponding TaxID and additional taxonomic information.

Subdivison	Species	TaxID	Additional information
Plantae	Marchantia polymorpha	3197	Marchantiophyta
Plantae	Apostasia shenzhenica	1088818	Tracheophyta
Plantae	Arabidopsis thaliana	3702	Tracheophyta
Plantae	Physcomitrella patens	3218	Bryophyta
Plantae	Chlorella variabilis	554065	Chlorophyta
Plantae	Cyanidioschyzon merolae	45157	Bangiophyceae
Unikonts	Homo sapiens	9606	Chordata
Unikonts	Lottia gigantea	225164	Lophotrochozoa
Unikonts	Apis cerana	7461	Ecdysozoa
Unikonts	Hydra vulgaris	6087	Cnidaria
Unikonts	Amphimedon queenslandica	400682	Porifera
Unikonts	Saccharomyces cerevisiae	4932	Dikarya
Unikonts	Catenaria anguillulae	109876	Blastocladiomycota
Unikonts	Gonapodya prolifera	1123529	Chytridiomycota
Unikonts	Monosiga brevicollis	81824	Choanoflagellates
Unikonts	Salpingoeca rosetta	946362	Choanoflagellates
Unikonts	Capsaspora owczarzaki	192875	Filasteria
Unikonts	Sphaeroforma arctica	72019	Ichthyosporea
Unikonts	Fonticula alba	691883	Nucleariidae
Unikonts	Dictyostelium discoideum	44689	Amoebozoa
Unikonts	Acanthamoeba castellanii	5755	Amoebozoa
Rhizaria	Plasmodiophora brassicae	37360	Cercozoa
Rhizaria	Reticulomyxa filosa	46433	Foraminifera
Chromalveolates	Vitrella brassicaformis	1169539	Alveolates
Chromalveolates	Toxoplasma gondii	5811	Alveolates
Chromalveolates	Paramecium tetraurelia	5888	Alveolates
Chromalveolates	Ectocarpus siliculosus	2880	Stramenopiles
Chromalveolates	Phytophthora parasitica	4792	Stramenopiles
Chromalveolates	Guillardia theta	55529	Hacrobia
Excavates	Leptomonas pyrrhocoris	157538	Euglenozoa
Excavates	Tritrichomonas foetus	1144522	Parabasalids

Excavates	Giardia intestinalis	5741	Fornicata
Excavates	Thecamonas trahens	529818	Apusozoa
Excavates	Naegleria gruberi	5762	Heterolobosea

Supplementary 2. List of query sequences used in BLAST search and orthologous sequences search in MetaPhors database. For each protein peroxiosmal ortholog from *Homo sapiens*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae* was selected.

Protein name	UniProtKB ID	Species	Query
Catalase	P04040	*Homo sapiens*	CTA1
Peroxisomal catalase A	P15202	*Saccharomyces cerevisiae*	CTA1
Catalase-2	P25819	*Arabidopsis thaliana*	CTA1
ATP-binding cassette sub-family D member 2	Q9UBJ2	*Homo sapiens*	PXA1
Peroxisomal long-chain fatty acid import protein 2	P41909	*Saccharomyces cerevisiae*	PXA1
ABC transporter D family member 1	Q94FB9	*Arabidopsis thaliana*	PXA1
ATP-binding cassette sub-family D member 1	P33897	*Homo sapiens*	PXA2
Peroxisomal long-chain fatty acid import protein 1	P34230	*Saccharomyces cerevisiae*	PXA2
Peroxisomal ABC transporter 1	F4JJ27	*Arabidopsis thaliana*	PXA2
Peroxisomal acyl-coenzyme A oxidase 1	Q15067	*Homo sapiens*	FOX1
Acyl-coenzyme A oxidase	P13711	*Saccharomyces cerevisiae*	FOX1
Peroxisomal acyl-coenzyme A oxidase 1	O65202	*Arabidopsis thaliana*	FOX1
Peroxisomal multifunctional enzyme type 2	P51659	*Homo sapiens*	FOX2
Peroxisomal hydratase-dehydrogenase-epimerase	Q02207	*Saccharomyces cerevisiae*	FOX2
Enoyl-CoA hydratase 2, peroxisomal	Q8VYI3	*Arabidopsis thaliana*	FOX2
Long-chain-fatty-acid--CoA ligase 1	P33121	*Homo sapiens*	FAA2
Long-chain-fatty-acid--CoA ligase 2	P39518	*Saccharomyces cerevisiae*	FAA2
Long chain acyl-CoA synthetase 7, peroxisomal	Q8LKS5	*Arabidopsis thaliana*	FAA2
Peroxisome biogenesis factor 1	O43933	*Homo sapiens*	PEX1
Peroxisomal ATPase PEX1	P24004	*Saccharomyces cerevisiae*	PEX1
Peroxisome biogenesis protein 1	Q9FNP1	*Arabidopsis thaliana*	PEX1
Peroxisome biogenesis factor 2	P28328	*Homo sapiens*	PEX2
Peroxisome biogenesis factor 2	P32800	*Saccharomyces cerevisiae*	PEX2
Peroxisome biogenesis factor 2	Q9CA86	*Arabidopsis thaliana*	PEX2
Ubiquitin-conjugating enzyme E2 D2	P62837	*Homo sapiens*	PEX4
Ubiquitin-conjugating enzyme E2-21 kDa	P29340	*Saccharomyces cerevisiae*	PEX4
Protein PEROXIN-4	Q8LGF7	*Arabidopsis thaliana*	PEX4
Peroxisomal targeting signal 1 receptor	P50542	*Homo sapiens*	PEX5
Peroxisomal targeting signal receptor	P35056	*Saccharomyces cerevisiae*	PEX5
Peroxisome biogenesis protein 5	Q9FMA3	*Arabidopsis thaliana*	PEX5
Peroxisome biogenesis factor 10	O60683	*Homo sapiens*	PEX10
Peroxisome biogenesis factor 10	Q05568	*Saccharomyces cerevisiae*	PEX10
Peroxisome biogenesis factor 10	Q9SYU4	*Arabidopsis thaliana*	PEX10
Peroxisomal membrane protein PEX14	O75381	*Homo sapiens*	PEX14
Peroxisomal membrane protein PEX14	P53112	*Saccharomyces cerevisiae*	PEX14
Peroxisomal membrane protein PEX14	Q9FXT6	*Arabidopsis thaliana*	PEX14

8. Curriculum vitae

Education:

2016-2018: Master of Molecular Biology, University of Zagreb, Croatia

- Top 10% of the class during the enrollment process
- Internship at the Ruđer Bošković Institute, Zagreb 1/2017-6/2017
- Internship at the CRG-Center for Genomic Regulation, Barcelona 3/2018-9/2018
- Master thesis: Evolution of core peroxisomal proteins –
  Advisors: Prof. Toni Gabaldón, PhD and Assoc. Prof. Damjan Franjević, PhD

2013-2016: Bachelor of Molecular Biology, University of Zagreb, Croatia

- Erasmus + student mobility program: Universitat de Vic-Universitat Central de Catalunya 1/2016-6/2016
- Student teaching assistant at Zoology course, Divison of Zoology, Department of Biology, Faculty of Science, University of Zagreb 10/2014-2/2015
- Bachelor thesis: Human genetic variability –
  Advisor: Assoc. Prof. Damjan Franjević, PhD

2009-2013: Classical Gymnasium, Zagreb, Croatia

Adwards/Fellowships:

- Rector's prize for individual scientific and artistic work awarded by the University of Zagreb

For scientific paper "Phylostratigraphic analysis of *Escherichia coli*" in the academic year 2016./2017. - Advisors: Assoc. Prof. Tomislav Domazet-Lošo, PhD and Assoc. Prof. Damjan Franjević, PhD

- Scholarship for Academic Excellence

Scholarship of the University of Zagreb for the most successful students in the academic year 2016./2017.

- Scholarship of the City of Zagreb for Academic Excellence

Scholarship given by the City of Zagreb for the most successful students in the academic year 2017./2018.

- Erasmus+ fellowship

For an exchange semestar at the University of Vic and for an internship at the CRG-Centre for Genomic Regulation

Conferences:

- ISCB Student Council Symposium 2017

Poster presentation "Phylostratigraphic analysis of *Escherichia coli*"

- Students' Symposium in Biology and Life Science 2017

Lecture "Phylostratigraphic analysis of *Escherichia coli*"

Other projects:

- Bioinformatics student section in BIUS

Co-founding and leading the Bioinformatics student section 9/2015-9/2017

- Night of Biology 2014. & 2015.

Performing lectures and experiments

- Summer School of Science 2014.

Leader of swapshop "Bioinformatics - a revolution in science" at the S3 programme of the Summer School of Science in Požega,Croatia

Languages:

- English C1 & French B2