pyham: *a python package to analyse hierarchical orthologous groups in orthoXML*

pyham is a python library to handle orthoXML files containing Hierarchical Orthologous Groups (HOGs). The motivation is to facilitate the extraction of evolutionary information contained in HOGs about either specific gene families or in aggregate. Depending on the functions, the output is provided as python data structures, as interactive javascript visualisations, or as graphs.

What are **H**ierarchical **O**rthologous **G**roup**s** (**HOGs**)?

Definition

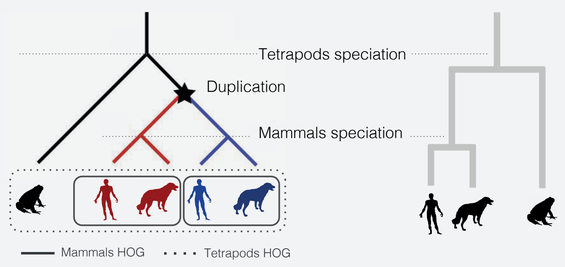
« **H**ierarchical **O**rthologous **G**roup**s** (**HOGs**) are defined as ***set of genes*** that all descended from a ***single common ancestral gene*** at a ***specific taxonomic range***. »

In terms of labeled phylogenetic gene trees, HOGs are corresponding to sub gene trees rooted by speciation events related to the HOG taxonomic range.

HierarchicalOrthologous Groups

Labeled gene tree (left) and its related specie tree (right) illustrating the evolutionary history of 5 genes all descended from a single common ancestral at the tetrapods level.

Those homologs can be classified as orthologs if they start diverging by speciation (blue genes, red genes) or as paralogs if they start diverging by duplication (blue and red genes).



We can identify in this example HOGs at two taxonomic levels: *one larger HOG at the tetrapods level* (dotted-line rectangle) containing all the homologous genes that emerged from the single tetrapod ancestral gene, and *two HOGs at the mammalian level* (solid-line rectangles), due to a duplication of the tetrapod ancestral gene before the mammals speciation.

Where to find them ?

HOGs inferred on public genomes can be found in the following orthology dabtase: **Eggnog** (eggnogdb.embl.de), **OrthoDb** (orthodb.org) and **OMA Browser** (omabrowser.org). If you want to use your custom genomes to infer HOGs you can use the **OMA Standalone Software** (omabrowser.org/standalone).

How does **pyham** help you investigate on HOGs?

Pyham working environment

Pyham takes as input an orthoXML file containing HOGs and the related species tree. Pyham creates *genomes and genes object* based on the given information and provide an API to work directly on those phylogenetic objects (*easy queries based on name or phylogenetic relations*). The inputted species tree serves as guideline to define genomes and genes evolutionary relationships.

EXAMPLE:
What is the evolutionary history of genes that happened in between ancestral genes of mammals and genes of humans?

Pyham provides a mapper object for HOGs/genes across multiple taxonomic ranges. The idea is to map the HOGs of an ancestral genome to the HOGs/genes of its descendant genome. The vertical mapper object allows to retreive all genes depending on their evolutionary history in between the two levels (which ones have duplicate?, which ones have been lost, etc...).

```
compare_human_mammals = pyham_analysis.compare_genomes_vertically('Human','Mammals')
compare_human_mammals.get_identical()  # Mammals HOGs with their single copy human descendant genes
compare_human_mammals.get_lost()  # Mammals HOGs that been lost in between the two levels
compare_human_mammals.get_gained()  # Human genes that have been "gained" in between the two levels
compare_human_mammals.get_duplicated()  # Mammals HOGs with their multiple copy human descendant genes
```

What are the genes in human that have duplicated in the branch leading to the speciation of mammals?

We can use logic operations in the previously described mapper object. In this case we can compare the genome of interest with its parent and retreive the duplicated genes that will be specific to this branch.

```
compare_mamm_tetra = pyham_analysis.compare_genomes_vertically('Mammals','Tetrapods')
mammals_specific_dupl_hogs = compare_human_mammals.get_duplicated()

human_genes_duplicated_before_mammals_speciation = []
for hog in mammals_specific_dupl_hogs:
    for gene in hog.get_descendant_genes()
        if gene.genome.name == 'Human':
            human_genes_duplicated_before_mammals_speciation.append(gene)
```

What is the number of ancestral genes in the Mammalian ancestral genome?

Ancestral genome objects act as proxy to fetch all hogs at specific taxon.

```
# return an ancestral genomes object
mammals_genome = pyham_analysis.get_ancestral_genome_by_name('Mammals')
# get the list of hogs in this ancestral genome
number_ancestral_geness_mammals = len(mammals_genome.genes)
```

How to get pyham ?

Pyham is available as python package on the pypi server and is compatible python 2 and python 3. You can easily install via pip:

```
pip install pyham
```

You can check the official pyham website (lab.dessimoz.org/py`ham) for further information about how to use pyham, documentation and others related ressources.

How can I visualise the evolutionnary history of a gene family (HOG)?

Pyham embedds **hogvis** , a tool to visualise gene family evolutionary history. The aim is to provide an interactive way of visualising ancestral genes composition of gene family.

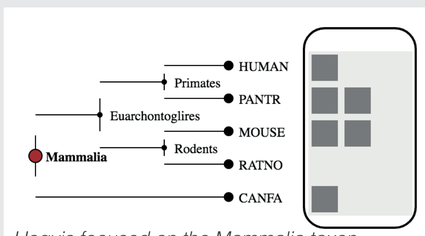
```
# Select an HOG
hog_of_interest =  pyham_analysis.get_hog_by_id(2)

# create and export the hog vis as .html
output_filename = "hogvis_example.html"
pyham_analysis.create_hog_visualisation(hog=hog_of_interest,outfile=output_filename)
```

As you can see in the figures aside, hogvis is composed of two panels: ***a species tree*** that allows you to select the taonomic range of interest, ***a genes panel*** where each grey square represents an extant gene and each line a species.

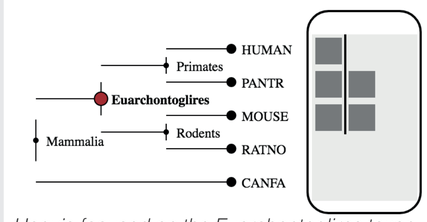
We can see for example in the top figure that at the level of mammals all genes of this gene family are descendant from a single comon ancestral gene.

If we are looking at the level of Euarchontoglires we observe that the genes are now split by a vertical line. This vertical line separates 2 group of genes that are each descendants from a same single ancestral genes. This is the result of a duplication in between Mammals and Euarchontoglires.

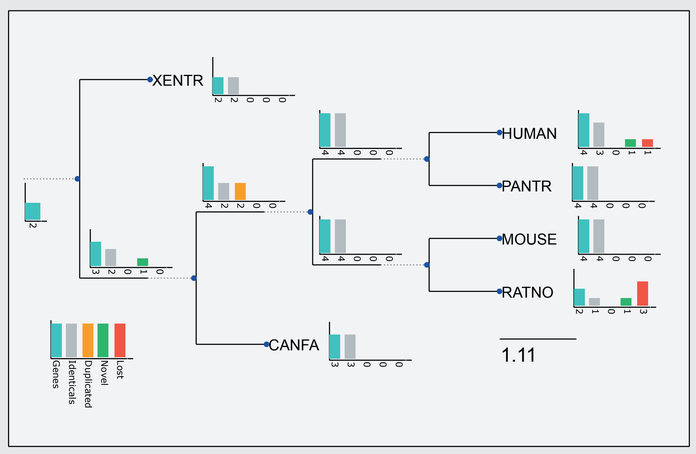
With a quick look we can easily identify for a gene family when duplication occured, which species have lost genes or how big genes families evolved.



Hogvis focused on the Mammalia taxon



Hogvis focused on the Euarchontoglires taxon



How can I visually represent the differents evolutionnary events that occured in my genomic setup?

Pyham includes **treeprofile** , a tool to visualise annotated species tree with evolutionary events (genes duplication, lost, gain) mapped to their related taxonomic range. The aim is to provide a minimalist and intuitive way to visualise the number of evolutionary events that occurred on each branches.

```
# create and export the treeprofile as .png (.svg, .pdf also available)
treeprofile=  pyham_analysis.create_tree_profile(outfile='example.png')
```

As you can see in the figure above, the treeprofile is composed of a the reference species used to perform the pyham analysis where each internal node is displayed with its related histogram of phylogenetic events (duplication, lost, gain, or no change) that occurred on each branch.

hogvis: gene family history visualisation

treeprofile: gains and losses on species tree

Clément-Marie Train[1,2,3], **Natasha M. Glover**[1,2,3], **Adrian M. Altenhoff**[1,2,3] and **Christophe Dessimoz**[1,2,3,4]

1 Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland
2 Swiss Institute of Bioinformatics, Lausanne, Switzerland
3 Center of Integrative Genomics, University of Lausanne, Lausanne, Switzerland
4 Department of Genetics, Evolution and Environment, University College London, London, UK