Cloud based system for storing, visualisation and analysis of genomics and metagenomics data of BIG4 Order

Presentation By :
Ashish Tomar

Project funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642241

Objectives

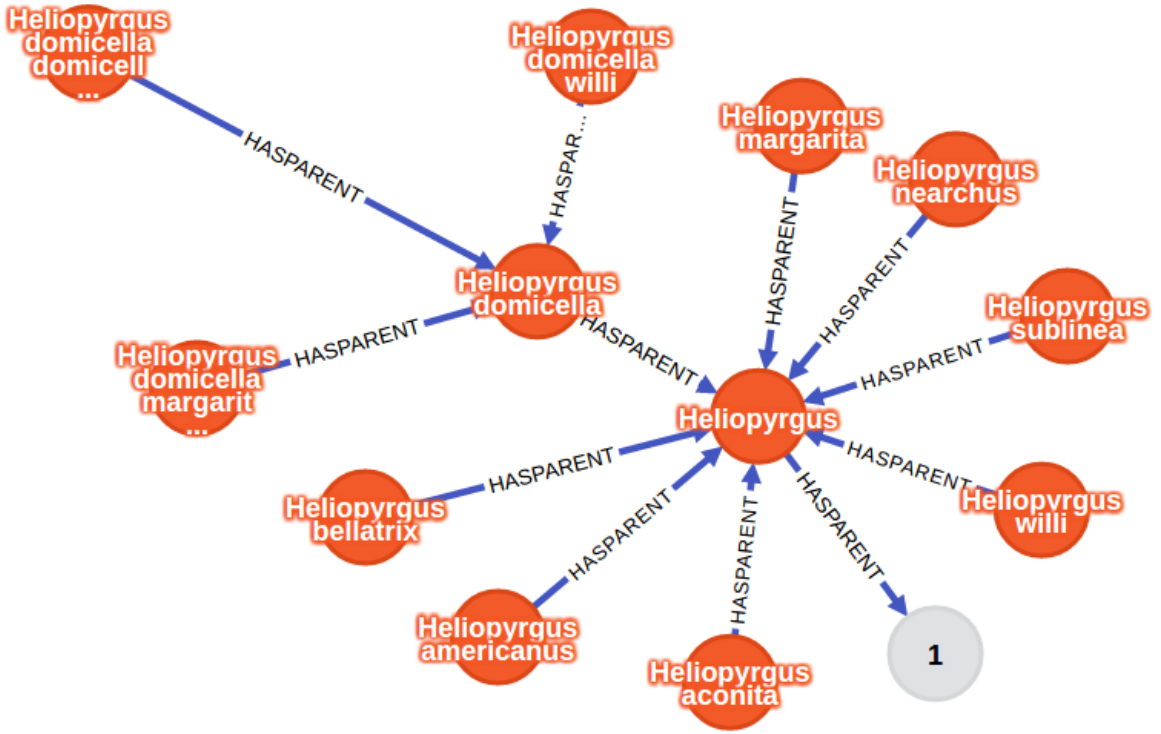
- Designing, developing and testing of innovative and interactive visualization tools for biological Big Data.
- To design, implement and test Bio4j queries and a biological interface adapted to the knowledge area of systematics.
- **Results**
  - One complete and two prototype applications for metagenomics visualization.
  - Phylogenetics Visualization library in Javascript
  - Phylogenomics visualization web app
  - Geographical phylogenetic data visualization prototypes

Graph Database

- A graph database is a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data.
- A key concept of the system is the graph (or edge or relationship), which directly relates data items in the store.
- The relationships allow data in the store to be linked together directly, and in most cases retrieved with a single operation.
- Graph databases employ nodes, edges and properties.
  - Nodes represent entities such as people, businesses, account, taxons etc.
  - Edges, also known as graphs or relationships, are the lines that connect nodes to other nodes; they represent the relationship between them (such as a node genus is parent of species or subspecies node)
  - Properties are pertinent information that relate to nodes. For instance a genus node having taxon id, parentId, taxon name, etc.

Neo4j And Taxonomy Data

- Representing taxonomy as graph data
- The butterfly genus Heliopyrgus having species and subspecies as child nodes.

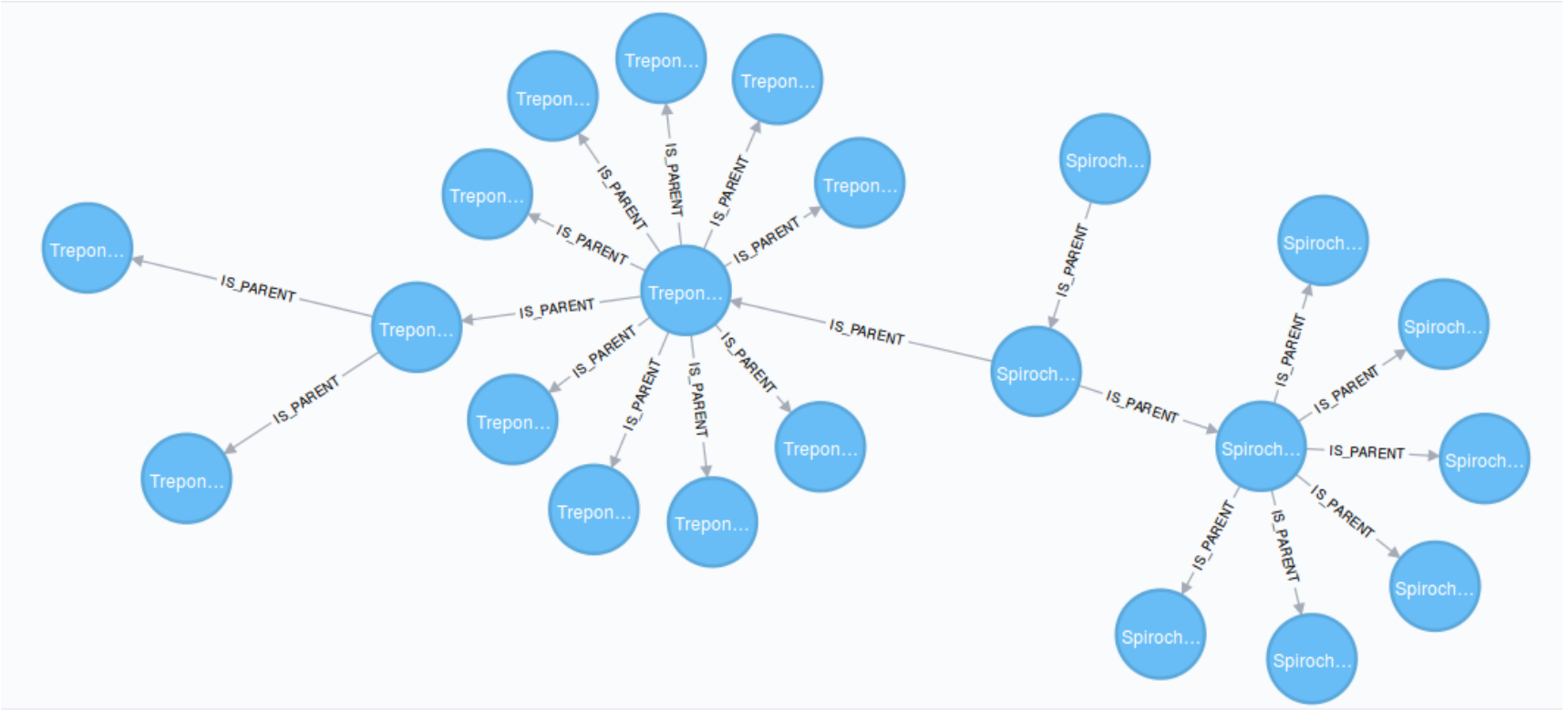


Taxonomy data in graph format.
Heliopyrgus genus is shown with sub species and species as the child nodes

Graph Database And NCBI Taxonomy Data

Bio4j is a bioinformatics graph based DB including

- UniProt (SwissProt + Trembl)
- Gene Ontology (GO)
- RefSeq
- NCBI taxonomy

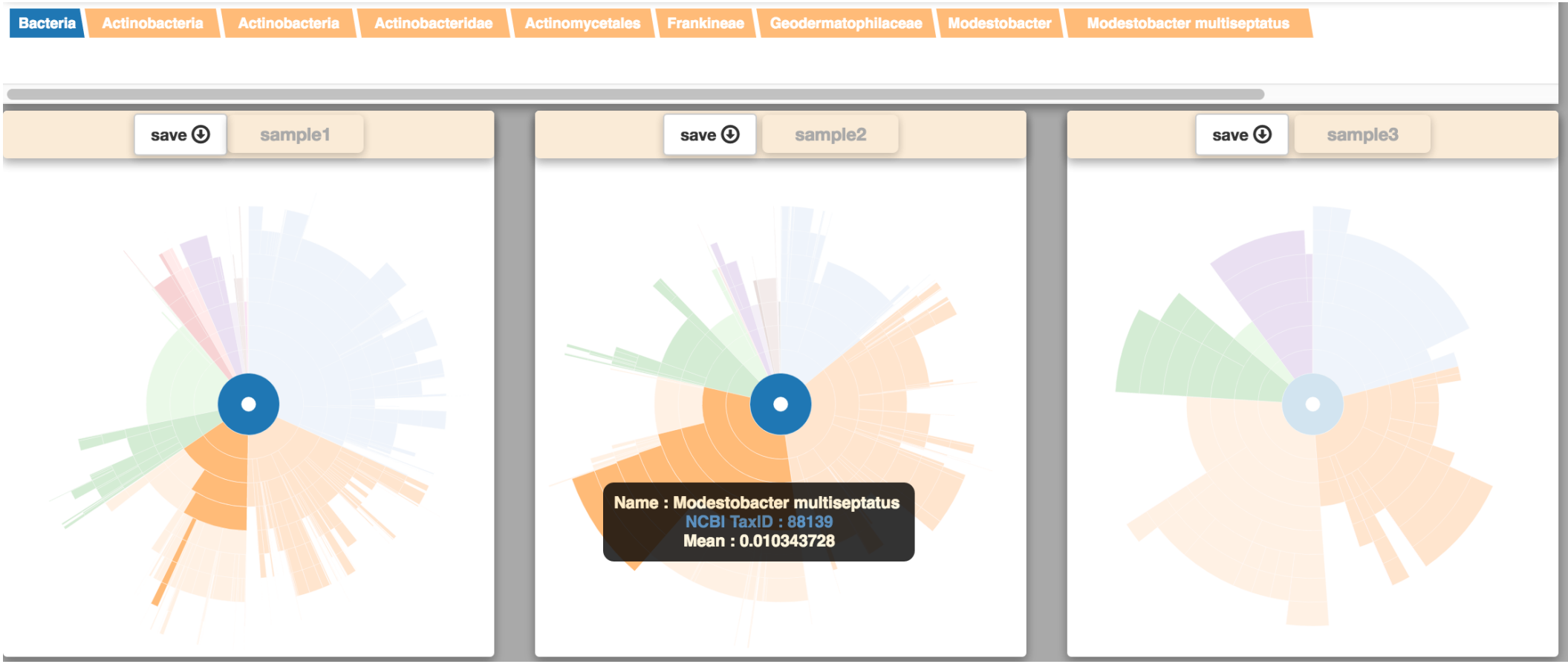


NCBI taxonomy data as graph database.

Metagenomics Data Visualization

- Distilling metagenomics data into graphical representations, however, is not a trivial task.
- The foundation of most metagenomics studies is the assignment of observed nucleic acids to taxonomic or functional hierarchies.
- Space filling diagrams (sunburst and treemap) convey hierarchy implicitly via angular subdivision.
- Furthermore, taxonomic and functional hierarchies are often too complex for all nodes to be shown, and wide variations in abundances can be difficult to represent.

Metagenomics Visualization Web Application



Phylogenetic Data Visualization

- D3.js has allowed rich and dynamic data visualization in bioinformatics web application.
- The exploratory analysis using the data integration and the visualization has made biological data more intuitive.
- Representation of tree data structures has become very popular in different biological fields like phylogenetics and metagenomics visualization
- Combining phylogenetic tree with other data such as expression data, protein sequence domains, gene structure etc. can make analysis more exploratory and intuitive.

Phylogenetic Data Visualization

- **Objective :**
  - A open source dynamic library for generating phylogenetic tree with genomics data integration based on d3.js
  - A Web based application interface for providing phylogeny tree visualization.
- **Library :**
  - D3.js : D3 is a JavaScript library for manipulating documents based on data.
  - Jquery : jQuery is a fast, small, and feature-rich JavaScript library.
  - Javascript : programming language of HTML and the Web
- **Input :**
  - Newick or nexus tree data format
  - Optional Data : Integration of genomic data such as expression data, gene structure, protein domain, etc. in form of JSON format

Phylogenetic Tree
Without Branch Scaling

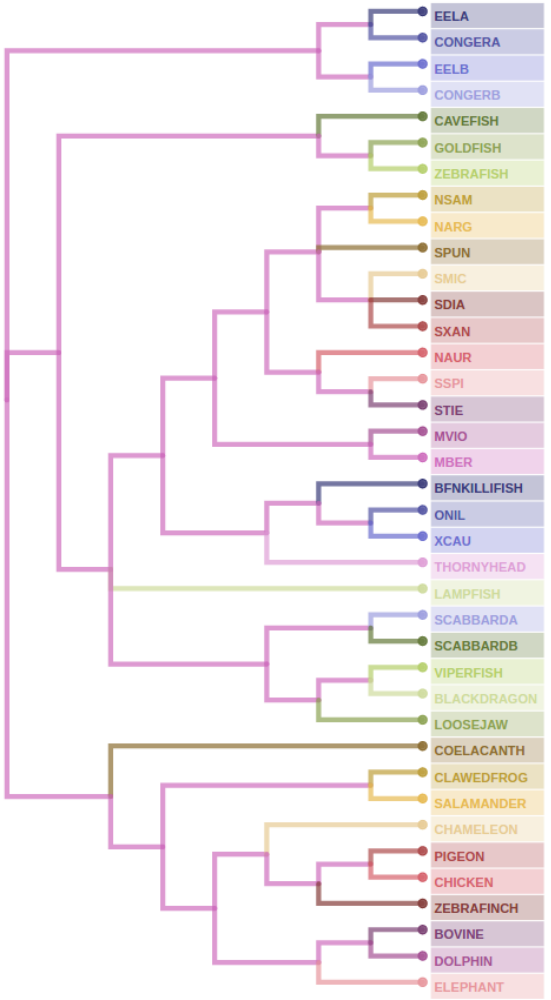


Fig1. Phylogenetic tree with no branch scaling

Phylogenetic Tree
Branch Length Scaling

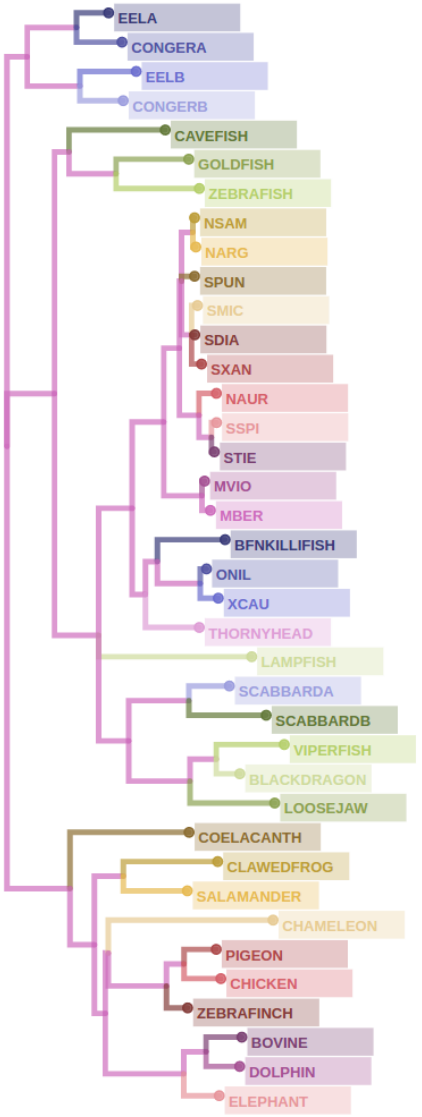


Fig2. Tree with branch scaling and no extended links

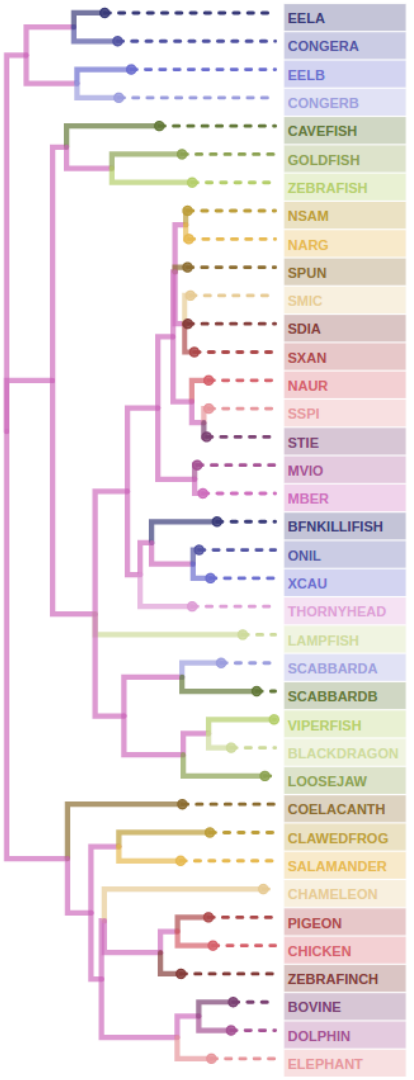


Fig3. Tree with branch scaling and extended links

Phylogenetic Tree And Bar chart

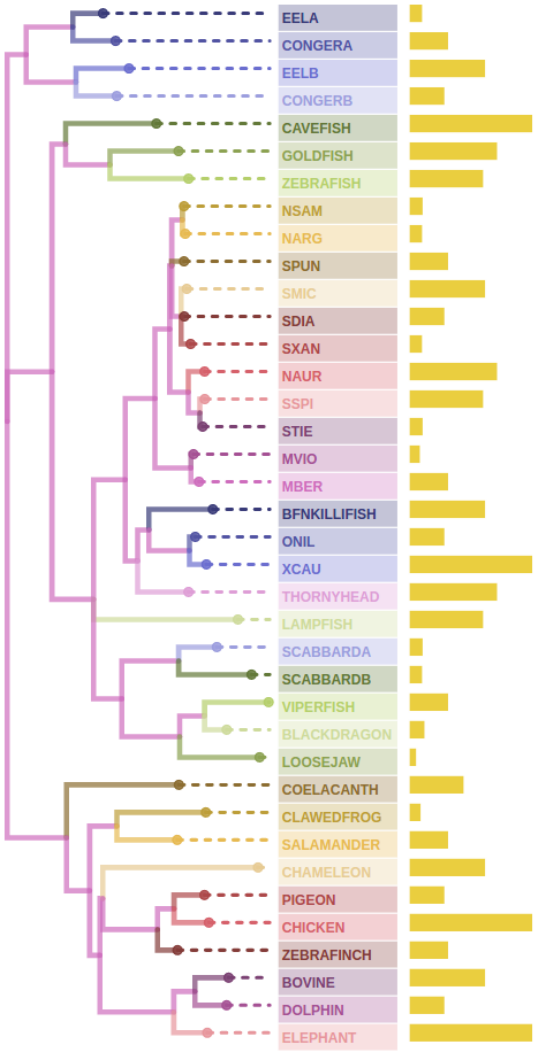


Fig4. Phylogenetic tree with bar chart

Phylogenetic Tree And Pie chart

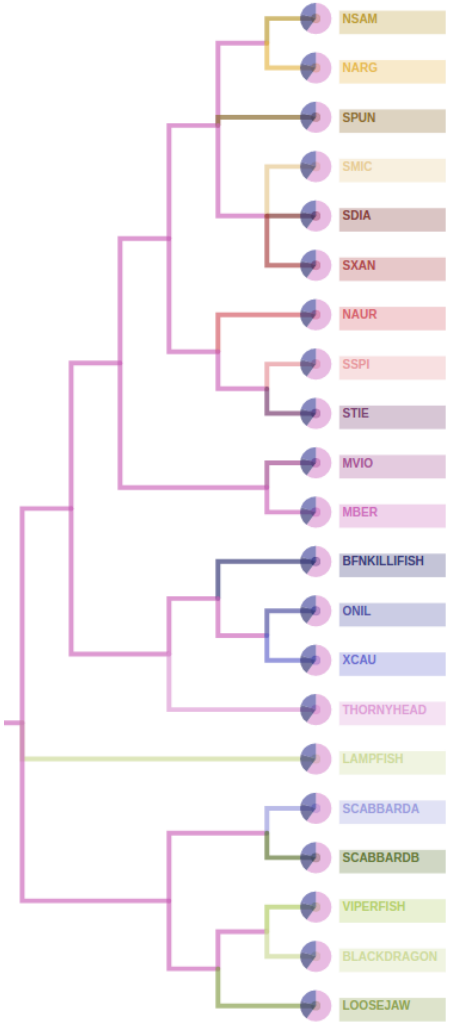


Fig 5. Phylogenetic tree with leaf nodes with pie chart

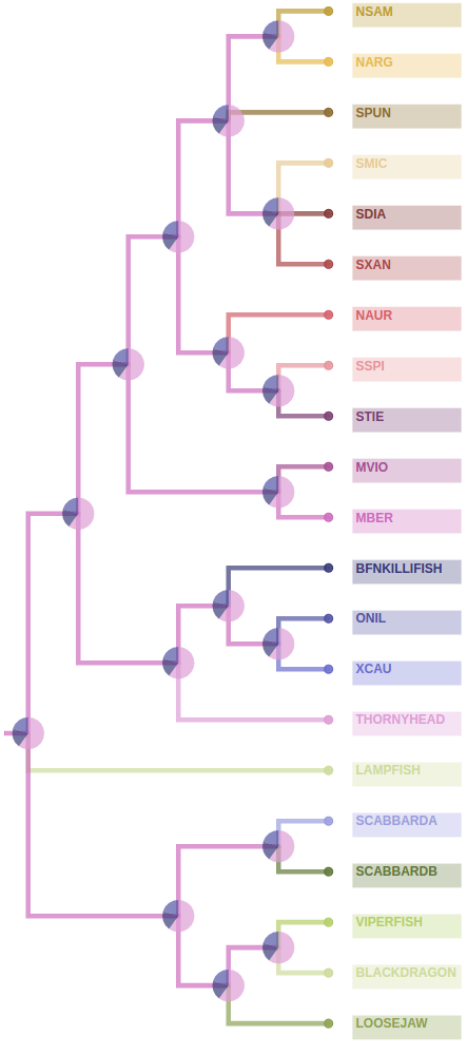


Fig 6. Tree with inner nodes with pie chart

Phylogenetic Tree And Sub tree

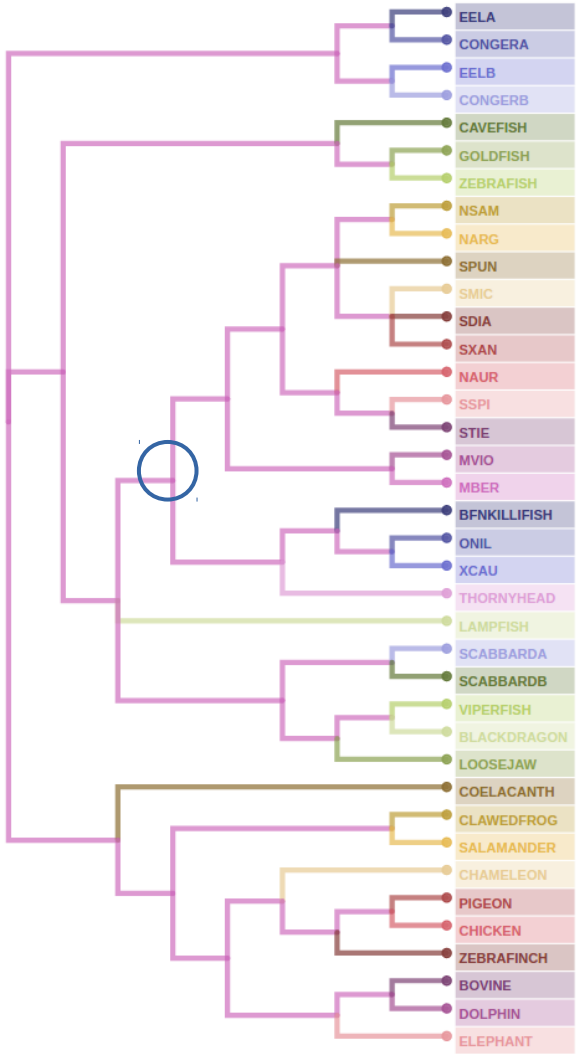


Fig 7. Pylogenetic tree (circle shows the branch point of subtree)

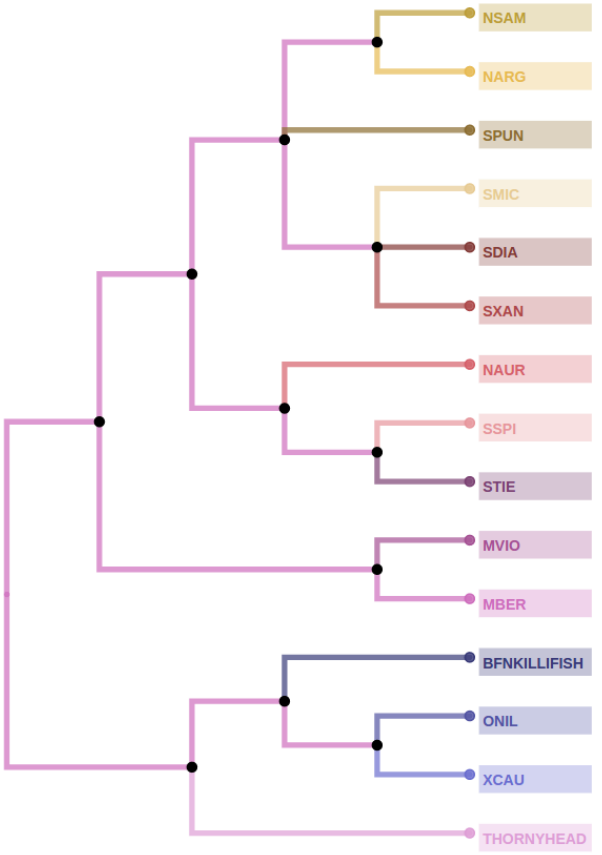


Fig 8. Sub tree after click on branch point of main tree

Phylogenetic Tree And Expression Data

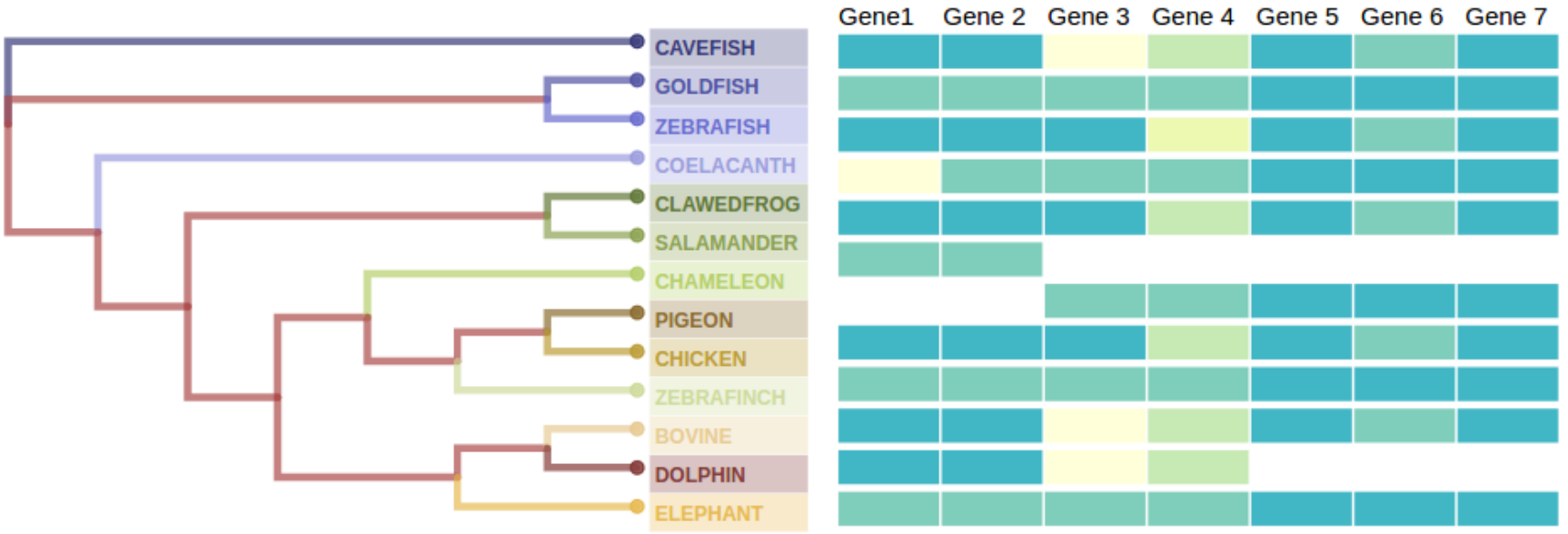


Fig 9. Phylogenetic tree with gene expression data

Phylogenetic Tree And Protein Domain Visualization

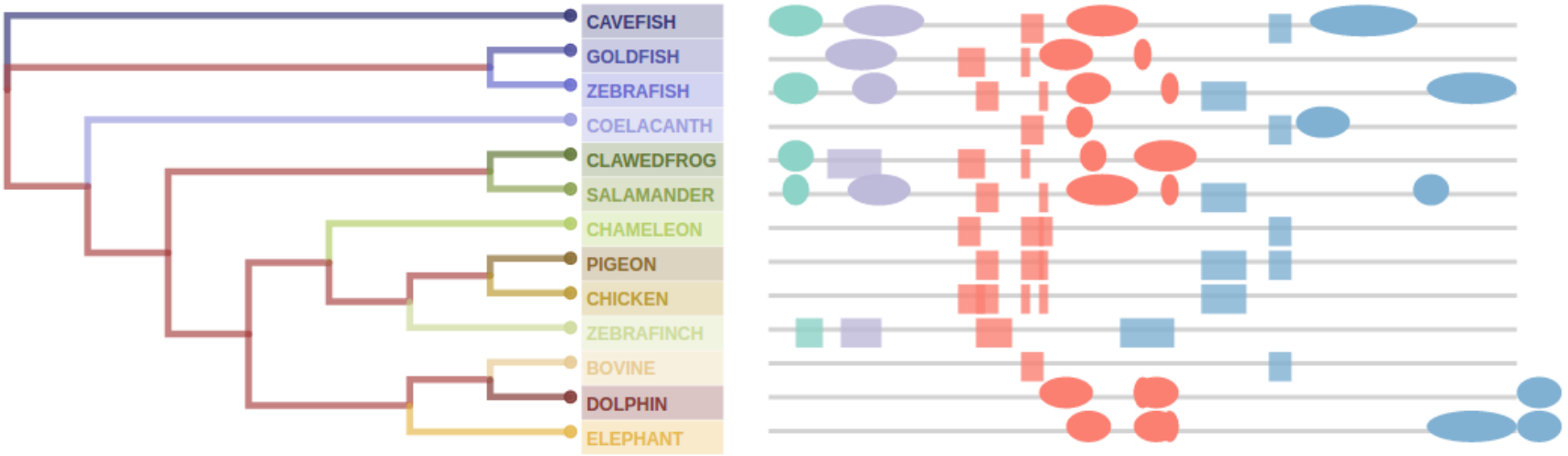


Fig 10. Phylogenetic tree with protein domain

Phylogenetic Tree And Gene Structure Visualization

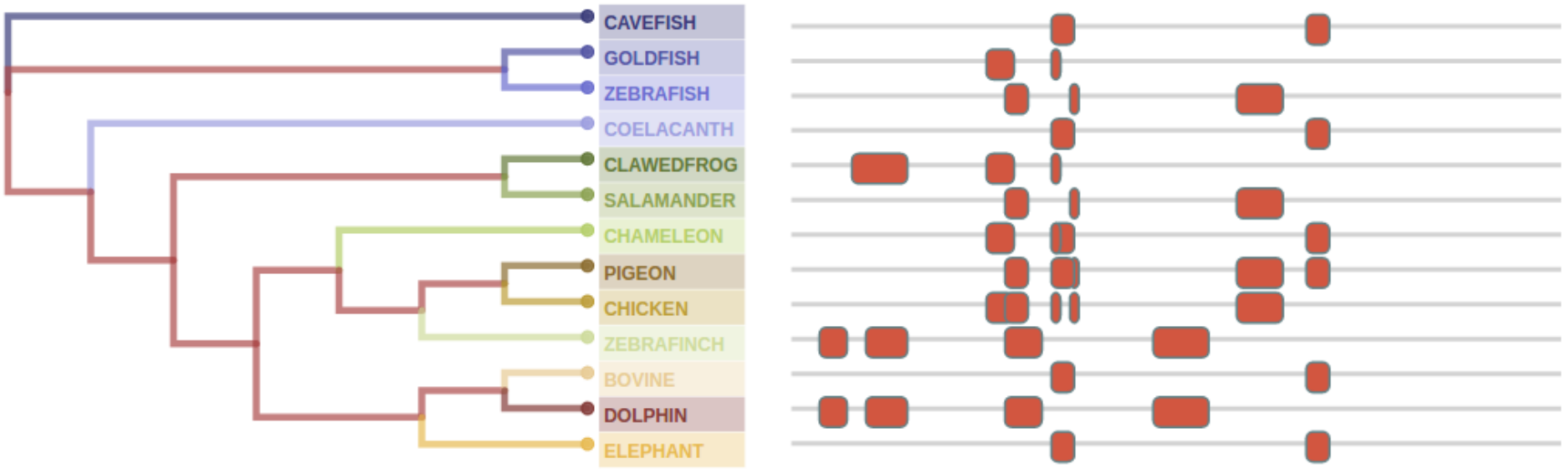


Fig 11. Phylogenetic tree with gene structure

Phylogenetic Radial Tree

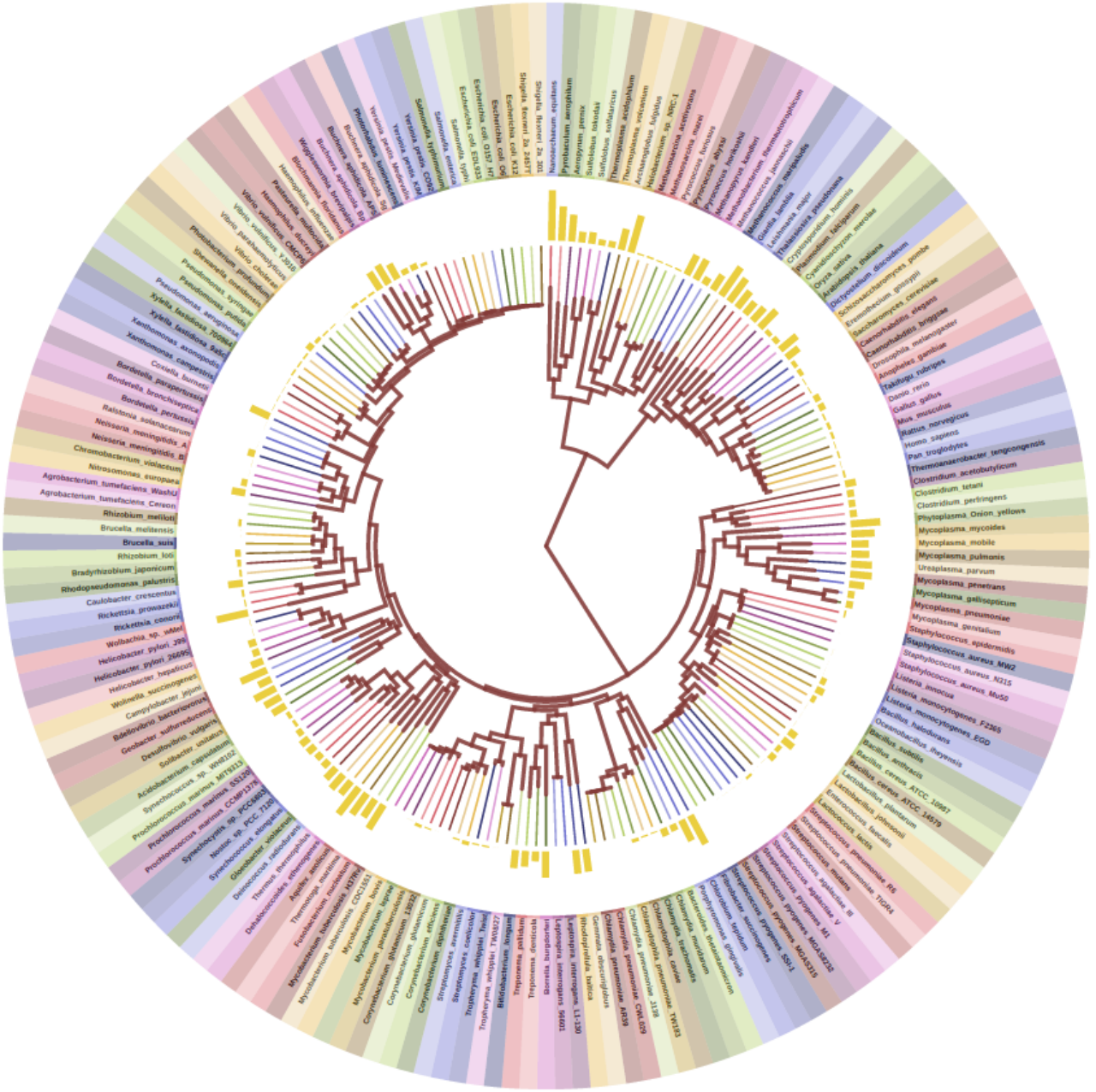


Fig 12. Phylogenetic radial tree

Geographical Phylogenetic Visualization

- Visualization of geographically distributed data will help in analyzing taxon wise species distribution.
- Geographically distributed data can be visualize with integration of other data such as IUCN status, heatmap, pie chart etc.
- Geographical visualization will help in understanding species richness or taxon data in different geographical location.
- Thus geographical visualization of “Big4 order” will increase the potential by simplifying the process of data exploration and analysis.

Geographical Phylogenetic Visualization

- Objectives :
  - To provide a framework to visualize and access information about the distribution of species and the BIG4 order composition geographically.
  - Visualize and explore aggregate diversity patterns such as species richness and other biogeographic patterns of different BIG4 order.
  - To create a framework to integrate data to biodiversity data for visual enrichment analysis.
  - To create a dynamic design for the visualization of biodiversity data

Geographical Phylogenetic Visualization

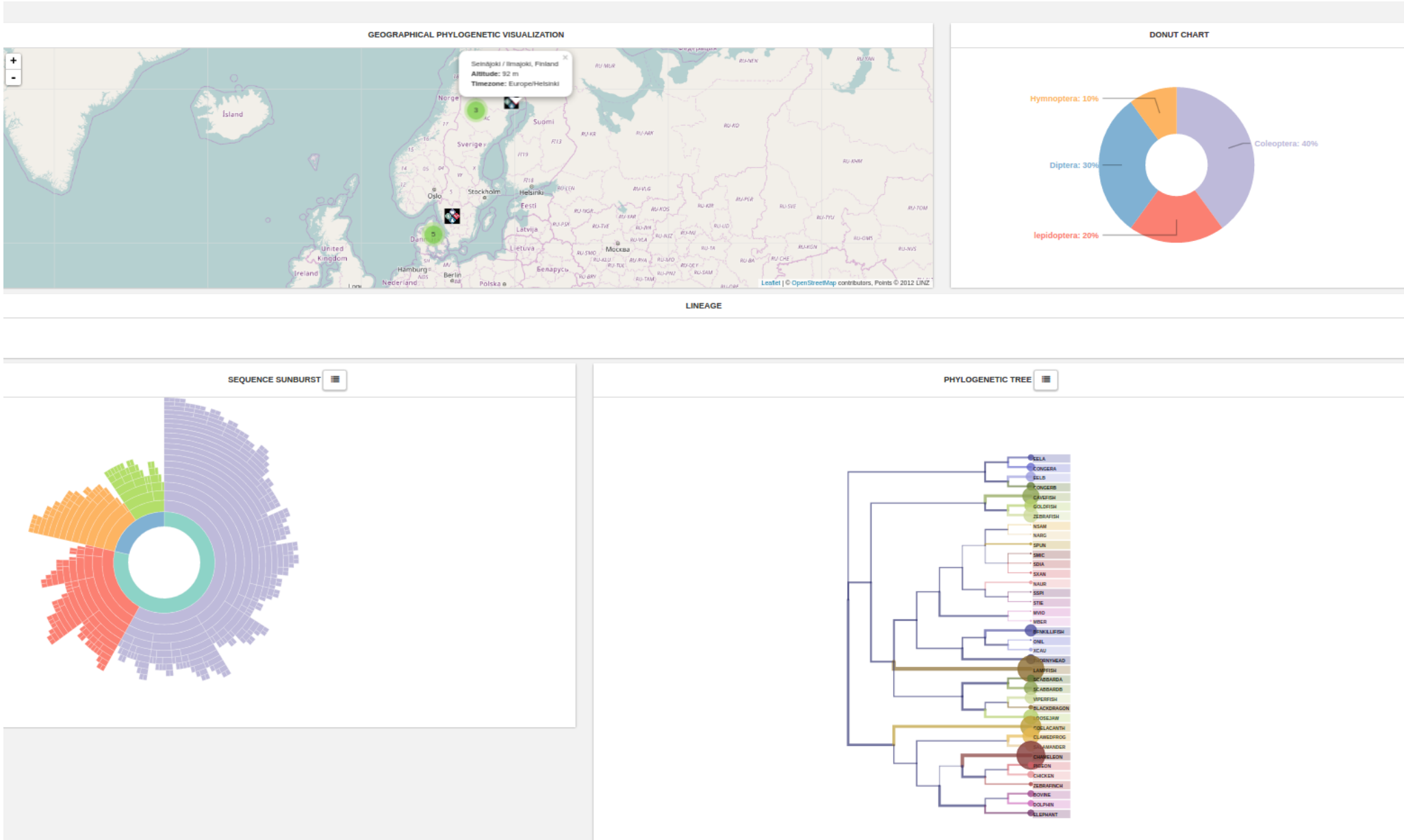


Fig 13. Phylogenetic tree information for clicked geographical location

Geographical Phylogenetic Visualization

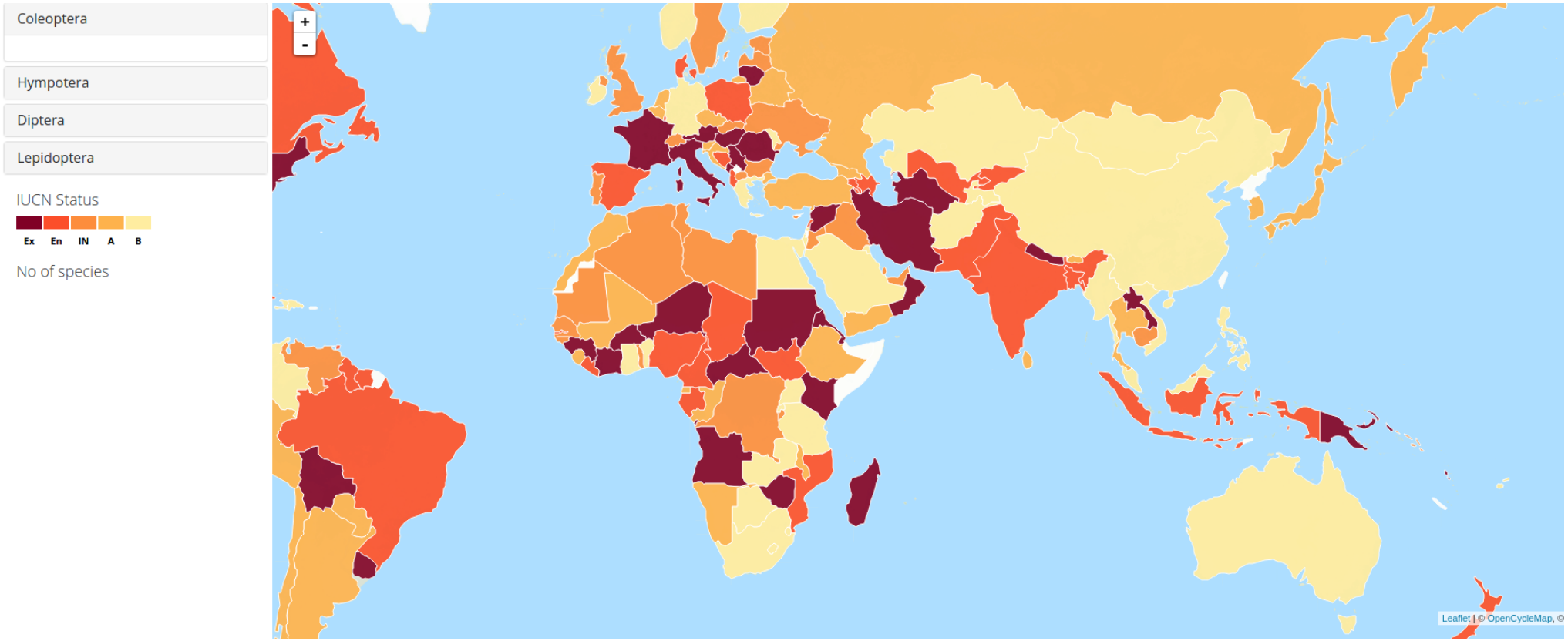


Fig 14. Phylogenetic tree information for clicked geographical location

Future Plans

- Creating REST api for querying bio4j based graph database for NCBI taxonomy database
- Metabarcoding visualization
- Upgrading features to geographical phylogenetic data visualization web application
- Develop a querying layer on top of Bio4j platform specific for BIG4 knowledge area at server side and a visualization layer at client side

Future Plans (PhD Thesis)
Tree Visualization in Biology

- A tree as a case of a graph and specially of a DAG (Directed Acyclic Graph).
- Identification of special problems dealing with large data sets as the whole taxonomy.
- Javascript, CSS , HTML5 and webGL
- Libraries: D3, Three.js, create.js and Sigma.js
- In summary, the objectives are:
  - To compile and analyze tree visualization tools in general.
  - To develop new methods and integrating them with third party solutions to solve some of these problems

Fundings

Funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642241

Thank you

References

1. http://neo4j.com/graphgist/e1552e40451ac0474548/

2. https://d3js.org/

3. www.w3schools.com/js/