Format for Efficient Storage of Homology Relations

Week 4 Report: Preliminary Experiment Using ETE3 and recPhyloXML

Kevin Gao

University of Toronto

June 23, 2022

Outlines

1. Parsing XML

2. References

ETE Toolkit

ETE is an open source Python library for tree analysis.

http://etetoolkit.org/

The library supports phyloXML and Newick format and provides API for accessing parsed trees: `ete3.phylo.Phyloxml` and `ete3.phylo.PhyloxmlTree`.

The underlying implementation uses `lxml`, which uses the DOM model, where the entire tree is parsed into the memory, without any caching or indexing.

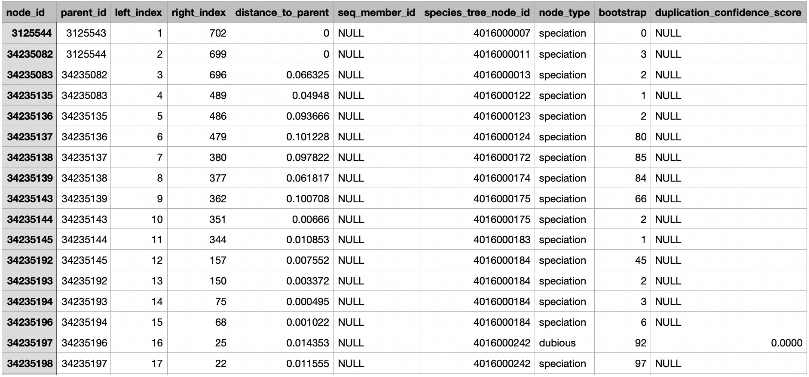
Another commonly used package, `BioPython`, uses `iterparse()`, which uses a model more similar to SAX (event-based parsing).

GeneTree API

We have implemented a set of APIs for parsing and interacting with gene trees encoded in phyloXML format (can be later modified to support recPhyloXML, if needed).

GeneTree API

We start from a plain phyloXML file exported from Ensembl Compara. We use the tree ID to query the root node in the mySQL database. The SQL query for obtaining the table is included in the repo.



node_id	parent_id	left_index	right_index	distance_to_parent	seq_member_id	species_tree_node_id	node_type	bootstrap	duplication_confidence_score
3125544	3125543	1	702	0	NULL	4016000007	speciation	0	NULL
34235082	3125544	2	699	0	NULL	4016000011	speciation	3	NULL
34235083	34235082	3	696	0.066325	NULL	4016000013	speciation	2	NULL
34235135	34235083	4	489	0.04948	NULL	4016000122	speciation	1	NULL
34235136	34235135	5	486	0.093666	NULL	4016000123	speciation	2	NULL
34235137	34235136	6	479	0.101228	NULL	4016000124	speciation	80	NULL
34235138	34235137	7	380	0.097822	NULL	4016000172	speciation	85	NULL
34235139	34235138	8	377	0.061817	NULL	4016000174	speciation	84	NULL
34235143	34235139	9	362	0.100708	NULL	4016000175	speciation	66	NULL
34235144	34235143	10	351	0.00666	NULL	4016000175	speciation	2	NULL
34235145	34235144	11	344	0.010853	NULL	4016000183	speciation	1	NULL
34235192	34235145	12	157	0.007552	NULL	4016000184	speciation	45	NULL
34235193	34235192	13	150	0.003372	NULL	4016000184	speciation	2	NULL
34235194	34235193	14	75	0.000495	NULL	4016000184	speciation	3	NULL
34235196	34235194	15	68	0.001022	NULL	4016000184	speciation	6	NULL
34235197	34235196	16	25	0.014353	NULL	4016000242	dubious	92	0.0000
34235198	34235197	17	22	0.011555	NULL	4016000242	speciation	97	NULL

GeneTree API

Once we have the reference table pulled from the database, we can annotate our gene tree (now parsed from phyloXML) to include duplication and speciation events, as well as the confidence score.

```
gt = GeneTree()
gt.load_phylo_xml(’test/test_data/gene_tree.xml’)
gt.load_ref_table(’test/test_data/ref_table.tsv’)
gt.annotate_event_nodes()
```

Export annotated phyloXML using `gt.export_phylo_xml(..)`.

Homology inference is done by inspecting the annotation on the lower common ancestor given two leaves.

GeneTree API

We plan to include a few benchmark testings for our primitive parser/API implemented in Python using ete. This will be our baseline benchmark. We will compare the result on larger trees with SAX-based parser (especially in terms of memory usage).

The short-term goal is to implement a phyloXML/recPhyloXML parser using the VTD-XML model with VTD and LC index. However, the library for VTD-XML parser was quite old and was not very well documented. It might take some time to document the parser code and get it to work with phyloXML.

Outlines

1 Parsing XML

2 References

T. C. Lam, J. J. Ding and J. Liu, "XML Document Parsing: Operational and Performance Characteristics," in Computer, vol. 41, no. 9, pp. 30-37, Sept. 2008, doi: 10.1109/MC.2008.403.

Haim Kaplan, Tova Milo, and Ronen Shabo. 2002. A comparison of labeling schemes for ancestor queries. In Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms (SODA '02). Society for Industrial and Applied Mathematics, USA, 954-963.

L. Nakhleh, D. Miranker and F. Barbancon, "Requirements of phylogenetic databases," Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings., 2003, pp. 141-148, doi: 10.1109/BIBE.2003.1188940.

Cardona, G., Rossello, F. and Valiente, G. Extended Newick: it is time for a standard representation of phylogenetic networks. BMC Bioinformatics 9, 532 (2008). https://doi.org/10.1186/1471-2105-9-532

Kmettlca, E.A. O(log n) persistent online lowest common ancestor search without preprocessing. https://github.com/ekmett/lca/

Wansong Zhang, Daxin Liu and Jian Li, "An encoding scheme for indexing XML data," IEEE International Conference on e-Technology, e-Commerce and e-Service, 2004. EEE '04. 2004, 2004, pp. 525-528, doi: 10.1109/EEE.2004.1287357.