Lineage Path Integration for Phylogenetic Resources

Katherine G. Herbert* Shashikanth Puspati† Jason T. L. Wang‡ William H. Piel§

Abstract

With the increase of genome and proteome data, phylogenetic information and phylogenetic analysis tools are increasing greatly in current biological repositories. First, many repositories allow users to browse information about species through taxonomic tools. These tools present the species with its lineage path and links to the various types of information the repository provides about the species. Second, some multiple sequence alignment tools offer users basic phylogenetic data through applying basic reconstruction algorithms to the alignment. With the availability of this information in multiple locations, integrated tools are needed to allow the user to compare this data. This paper presents data integration research on lineage paths using the BIO-AJAX framework. It introduces BIO-AJAX for Lineage Paths, a tool that integrates lineage path information for NCBI Taxonomy Database [1] and the Integrated Taxonomic Information System (ITIS)[6].

1. Introduction

Biological data, specifically phylogenetic data, is rich with issues that can be addressed with data quality and integration methodologies. Data quality in biological data is an important function necessary for the analysis of biological data. It can standardize the data for further computation and improve the quality of the data for searching. Data integration is also an important function necessary for analyzing biological data [3, 5, 7, 8, 9]. The very core purpose for most biological databases is to create a repository, integrating work from numerous scientists. Phylogenetic data encapsulates these problems. It is a complex data set that consists of processed wet lab results, data obtained through knowledge discovery, structural data and metadata. Moreover, phylogenetic studies also have multiple applications from Tree of Life studies to biomedical investigations [14].

Lineage path studies represent a unique opportunity in phylogenetic data integration. Currently, many repositories model lineage paths in some manner. These lineage paths can represent various types of knowledge including the taxonomy of a species, the development of the Tree of Life with respect to a databases data model or the intermediary nodes developed from a phylogenetic construction. With these different purposes for lineage paths, finding similarity between repositories concerning lineage paths can be confusing and sometimes inconsistent from the user's point of view.

This short paper introduces research issues and concerns regarding the lineage path integration and data quality problems in phylogenetic studies. Lineage paths themselves tend to relate to taxonomy studies. However, since this method can be employed on both taxonomic databases and phylogenetic tree databases, we will discuss this tool in reference to its phylogenetic purposes. It first identifies the lineage path problem, discussing why it is of interest to researchers. Next, it discusses applying data quality and data integration techniques to lineage paths to create an environment that facilitates user comparisons of lineage paths. Finally, it identifies future work for this project as well as makes some concluding remarks.

2. Lineage Paths

A lineage path is the path from a given point on a phylogenetic tree(such as the Tree of Life) to a specific taxon. Sometimes this taxon can be a species, which tends to be a terminal node on the Tree of Life or it can be an intermediary node within the tree. Most lineage paths concern the path from the root node of the Tree of Life to a specific taxon. Lineage paths pose many different problems for phylogenetic researchers. For example, in phylogenetic nomenclature, the lineage or ranking of an evolutionary unit or taxon is not standard. NCBI Taxonomy Database [1] uses 28 distinct ranks for classification while the International Code of Botanical Nomenclature uses 25 rankings [13]. These databases maintain, besides different categories

[*] To whom correspondence should be addressed: Department of Computer Science, Montclair State University, Montclair, NJ 07043, USA, herbertk@mail.montclair.edu

[†] Department of Computer Science, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA.

[‡] Department of Computer Science, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA.

[§] Department of Biological Sciences, 608 Cooke, University at Buffalo, Buffalo, NY 14260, USA.

for their taxonomies, depreciated lineage paths. Therefore, semantic integration must be used so that the matching of the rankings and the treatment of the depreciated paths are correct. When adding the complexity of a research phylogenetic tree repository such as TreeBASE, where there are multiple trees for one species and non-standard nomenclature, integration becomes more complex.

Lineage paths and lineage path comparison offers phylogenetic researches many interesting functionalities for their research. Through their comparison, researchers can analyze differences in phylogenetic reconstruction methods, understand differences in rankings among phylogenetic databases as well as perform some advanced queries upon the phylogenetic studies. Some possible queries of interest associated with lineage paths include: Compare the lineage paths for taxon X from database D1 and database D2; Given taxon X, find all lineage paths containing taxon X and; Given taxon X, find all taxa which are descendents (or ancestors) of X.

Foremost, lineage path querying offers the ability to compare ranking systems among the supported databases. For example, as mentioned previously, there are differences in ranking systems from database to database. By offering lineage path querying, a user can compare side by side the differences between the paths among the different databases. These rankings are important since the scientific name of the species is dependent upon where it falls within a given ranking system. If a species is classified differently from database to database, it affects the type of data that can be retrieved about a particular taxon. A user may have an understanding of one ranking system without understanding another. By allowing for comparison, we permit the user to evaluate each databases ranking system, seeing where the data he is interested in may be located [11, 13].

Extending these path comparisons to phylogenetic studies, it can also help scrutinize the differences between two trees which have similar taxa but use different reconstruction methods. By breaking the trees down to their paths, we can monitor the differences in the classification of a specific taxon through various reconstruction methods. This can be useful in analyzing the effectiveness of a reconstruction method as well as determining a species proper ranking.

Finally, lineage path queries can also help with advanced querying. One extremely powerful query that no phylogenetic database supports effectively is given a node N of the phylogenetic tree, find all ancestors or descendents from N. To perform this type of search, most databases require that the user navigate through some hierarchical browsing method of analyzing the tree. To find a specific path of the tree, the user must be familiar with the tree. For example, if a novice user tries to find “Homo sapiens” from the root of the tree, most novices would not understand enough about phylogenetics to know that the first classification “Homo sapiens” falls under is “Eukaryota” . Therefore, a user can enter a portion of the path, for example “Homo sapiens”, and get the path beginning (or ending) at “Homo sapiens”. Also, if the user is unclear about what type of organisms would fall under “Eukaryota”, he can enter “Eukaryota” as a partial path query, receiving back all paths that contain “Eukaryota”.

3. The BIO-AJAX Toolkit for Lineage Paths

The BIO-AJAX [4] toolkit facilitates improving the state of lineage path querying through integrating lineage path resources. By applying the data cleaning architecture employed by BIO-AJAX [2, 4], various lineage path resources can be integrated together to offer the user the ability to query and compare these lineage paths.

The BIO-AJAX framework for Lineage Paths has been implemented using the NCBI Taxonomy Database and the Integrated Taxonomic Information System (ITIS) [6]. Both tools allow querying upon their taxon set and return lineage paths as a part of the answer to the query. In the current implementation of BIO-AJAX for Lineage Paths, the following queries are addressed: Compare the lineage paths for taxon X from database D1 and database D2 and; Given taxon X, find all taxa which are descendents (or ancestors) of X.

In future versions of this tool, the option of finding any lineage path that contains taxon X will be provided. Moreover, other database repositories' lineage paths, such as TreeBASE's will also be incorporated into the integrated framework.

4. Implementation

The current implementation for BIO-AJAX for Lineage Paths uses the data warehousing technique for integrating the data and is accessible through a World Wide Web interface [5]. Figure 2 displays the interface for this tool. The lineage paths are extracted from both NCBI Taxonomy Database and ITIS and stored locally. In previous versions of the tool and previous instantiations of BIO-AJAX, the mediator method was used to display the paths. However, due to limitations in accessing each repository, this method had to be replaced with the warehousing method. Moreover, since each lineage path needed to be manipulated to get very specific data out of it, the mediator method became impractical. Also, for finding all ancestors and descendents, the lineage path strings needed to be parsed creating a long lag time for the Web interface. Therefore, the platform was shifted from using the mediator method to the data warehousing method. Figure 1 demonstrates the configuration of the system.

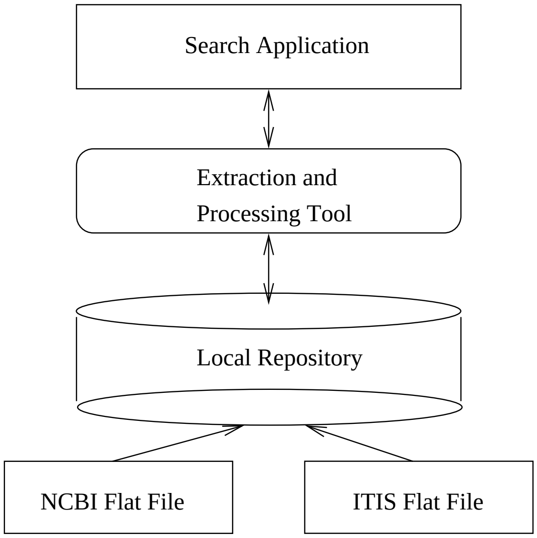


Figure 1. The software architecture of BIO-AJAX for Lineage Paths.

4.1. Lineage Path Extraction

For both data repositories, lineage paths can be extracted from the flat files each repository provides for download. While both databases vary in format, both store the lineage path information in similar ways. Both databases have signified one taxon as a "root" for the tree of life its data model represents. Each taxon in both repositories has various types of information stored about it, including the taxon id and the number of its immediate ancestor. Therefore, to obtain any taxons lineage path, a user would traverse the flat file recursively, until he obtains the root taxon.

Initially, the flat files from the repositories are downloaded to local storage in a MySQL database. Next, the scientific nomenclature is associated with its taxon identification number. To facilitate traversal of the paths, all taxa are also indexed through their taxon identification numbers. For both the ancestor and descendent tables, the paths are extracted from these flat files and stored locally. Moreover, the ranking files from each database are also extracted to provide the user with information about the terms in the paths so that informed comparison is possible.

4.2. Lineage Path Retrieval

Querying for the purpose of comparison, for finding all ancestors and for finding all descendents poses different problems that must be addressed so that the query is executed properly. Concerning querying for comparison and querying to find all ancestors, the data returned is very similar for both queries. However, for finding all descendents, the path must be manipulated in a different way to obtain these answers both efficiently and effectively.

To execute these three queries, a user can interact with the indices through a web based search mechanism. On this web page a user enters a taxon name within a text box. He next selects from three choices (Ancestor, Descendent and Home) listed beneath the text box what type of query he wants executed. From this, the appropriate query is executed. For each query, results from both NCBI and ITIS are displayed along with the ranking for each member taxon of the lineage path. Figure 2 displays the output for an ancestor query.

4.3. Lineage Paths and Data Cleaning

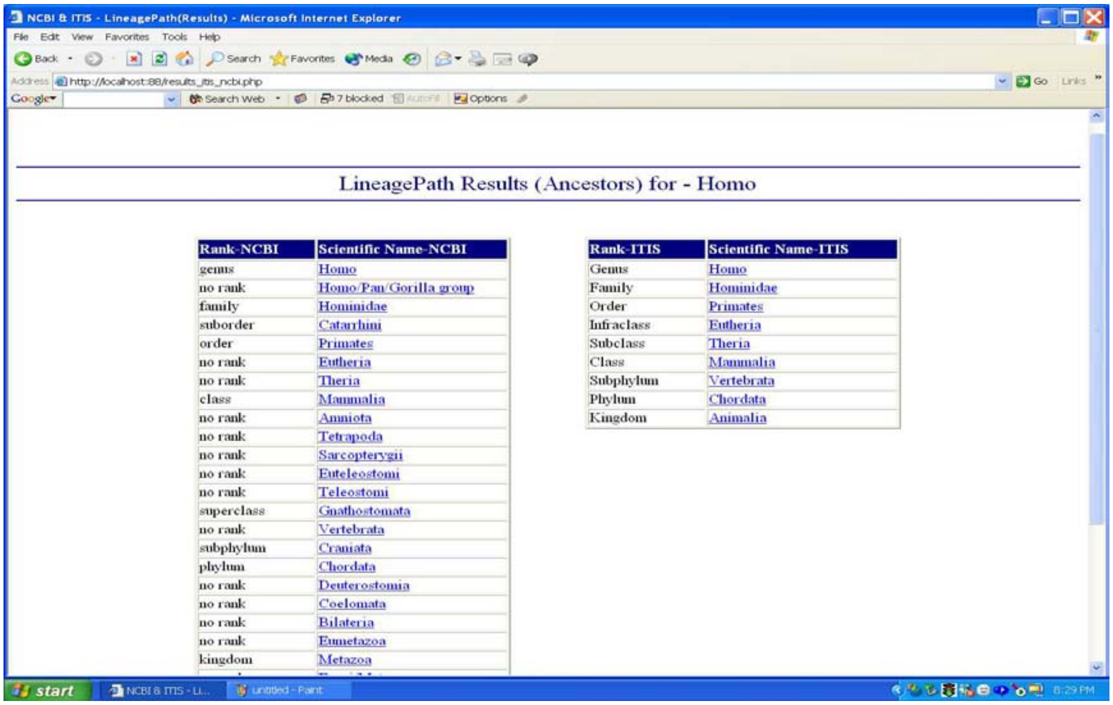
In the current method for finding the descendents, a number of taxa can possibly be returned numerous times, depending upon how close the query taxon is to the root of the tree when employing the descendent query. Therefore, data cleaning methods can be applied to these paths to eliminate redundancy [10, 15].

One possible method for eliminating redundancy is to apply the sorted neighborhood method to the array to detect similarities. Since the paths are highly structured and common errors such as spelling errors are rare, this becomes an ideal application for using the sorted neighborhood method. In a simple implementation of the sorted neighborhood method on lineage paths, the very least common paths between taxa on the same level of the tree can be found by comparing the lineage paths up to but not including the final taxon in the path. Identical paths can be merged so that redundancy is eliminated. For more advanced applications, there can be a number of iterations of this comparison, where the first iteration starts by comparing for one common taxon and builds to more complex paths.

5. Conclusion and Future Work

Lineage paths offer an interesting and rich method for phylogenetic researchers to explore various interpretations of the Tree of Life. By creating a comparative environment for phylogenetic researchers, problems concerning the correctness of the Tree of Life can be addressed.

Future work concerning this research includes integrating more repositories into the tool as well as improving the user interface [12, 16]. Moreover, this work can be further applied to consensus tree and supertree generation problems. Also, work can be done to further the visualization aspects of this tool concerning the comparisons.


NCBI & ITIS - LineagePath(Results) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://localhost:88/results_itis_ncbi.php
Go Links
Google Search Web 7 blocked AutoFill Options
start
NCBI & ITIS - Li...
untitled - Paint
8:29 PM

LineagePath Results (Ancestors) for - Homo

Rank-NCBI	Scientific Name-NCBI
genus	Homo
no rank	Homo/Pan/Gorilla group
family	Hominidae
suborder	Catarrhini
order	Primates
no rank	Eutheria
no rank	Theria
class	Mammalia
no rank	Amniota
no rank	Tetrapoda
no rank	Sarcopterygii
no rank	Euteleostomi
no rank	Teleostomi
superclass	Gnathostomata
no rank	Vertebrata
subphylum	Craniata
phylum	Chordata
no rank	Deuterostomia
no rank	Coelomata
no rank	Bilateria
no rank	Eumetazoa
kingdom	Metazoa

Rank-ITIS	Scientific Name-ITIS
Genus	Homo
Family	Hominidae
Order	Primates
Infraclass	Eutheria
Subclass	Theria
Class	Mammalia
Subphylum	Vertebrata
Phylum	Chordata
Kingdom	Animalia

Figure 2. The BIO-AJAX interface for showing an example output for an ancestor query.

References

[1] S. Federhen, I. Harrison, C. Hotton, D. Leipe, V. Soussov, R. Sternberg, and S. Turner. NCBI Taxonomy Homepage, `http://www.ncbi.nlm.nih.gov/ Taxonomy/taxonomyhome.html/`, 15 January 2005.

[2] H. Galahardas, D. Florescu, D. Shasha, E. Simon and C.A. Saita. Declarative Data Cleaning: Language, Model and Algorithms, in *Proc. of 27th International Conference on Very Large Data Bases*, September 11-14, 2001, Roma, Italy : 371-380.

[3] A.Y. Halevy. Answering queries using views: A survey. *Very Large Database Journal*, 10(4): 270-294, 2001.

[4] K.G. Herbert, N.H. Gehani, J.T.L. Wang, W.H. Piel and C.H. Wu. BIO-AJAX: An Extensible Framework for Biological Data Cleaning. *ACM SIGMOD Record*, 33(2): 51-57, June 2004.

[5] T. Hernandez and S. Kambhampati. Integration of Biological Sources: Current Systems and Challenges Ahead. *ACM SIGMOD Record*, 33(3):51-60, September 2004.

[6] The ITIS Organization, The Integrated Taxonomic Information System (ITIS) `http://www.itis.usda.gov/`, 15 January 2005.

[7] Z. Lacroix and T. Critchlow. *Bioinformatics: Managing Scientific Data*, Morgan Kaufmann, 2003.

[8] M. Lenzerini. Data integration: a theoretical perspective. In *Proc. of the 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002.

[9] B. Ludäscher, A. Gupta, and M.E. Martone. "A Model Based Mediator System for Scientific Data Management." Eds. Z. Lacroix and T. Critchlow, *Bioinformatics: Managing Scientific Data*, Morgan Kaufmann Publishers, 2003, pp. 335-370.

[10] W.L. Low, M.L. Lee, and T.W. Ling. "A knowledge-based approach for duplicate elimination in data cleaning." *Information Systems*, 26:8(Dec. 2001): 585-606.

[11] L. Nakhleh, D.P. Miranker, F. Barbancon, W.H. Piel and M. Donoghue. Requirements of Phylogenetic Databases, in *Proc. Of the 3rd IEEE International Symposium on BioInformatics and BioEngineering*, 10-12 March 2003, Bethesda, MD: 141-148.

[12] W.H. Piel, M.J. Donoghue, and M.J. Sanderson. Treebase: A database of phylogenetic information. In *Proceedings of the 2nd International Workshop of Species 2000*, 2000.

[13] R.D.M Page: Phyloinformatics: Towards a Phylogenetic Database, in *Data Mining in Bioinformatics*, Wang, J.T.L, et al., Eds., October 2004.

[14] R.D.M. Page and E.C. Holmes. *Molecular Evolution: A phylogenetic approach*. Blackwell Science, 1998.

[15] E. Rahm and H.H. Do. Data Cleaning: Problems and Current Approaches , in *Bulletin of the Technical Committee on Data Engineering, Special Issue on Data Cleaning*. 23.4 (Dec 2000): 3-13.

[16] M. J. Sanderson, M.J. Donoghue, W. H. Piel, and T. Eriksson. Treebase: A prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *Am. J. Bot.*, 81(6), 1994.