Phylogenetic Information Integration: Research Issues and Techniques

Katherine G. Herbert* and Jason T. L. Wang†

Abstract

Phylogenetic research is currently experiencing significant growth due to the generation of massive biological data sets. Moreover, many different disciplines are seeing the benefits in such research. However, due to the complexity of the data set, phylogenetic data has become disjoint. This paper discusses methods for integrating phylogenetic data. It looks at currently used practices and mentions future trends in phylogenetic data integration. Finally, it examines how data quality can help with this integration.

1 Introduction

Research concerning phylogenetic data and its applications is growing at a signi cant rate. Due to the explosion of genomic and proteomic data, more and more research is turning towards phylogenetic studies. Moreover, as computer models are demonstrating the use and power of hierarchical studies, more elds outside of the traditional phylogenetic studies are using phylogenetic models to discover information.

With this explosion of interest in phylogenetic information and modeling, a diaspora is developing between various categories of research within phylogenetic studies. For the user, he or she usually can only access information about speci c methodologies of study. For example, there are numerous repositories that address lineage path and taxon information, such as the NCBI Taxonomy Database (`http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy`), the Integrated Taxonomic Information System (`http://www.itis.usda.gov/`), the Glasgow Taxonomic Name server (`http://darwin.zoology.gla.ac.uk/~rpage/MyToL/www/index.php`), CladeStore (`http://palaeo.gly.bris.ac.uk/cladestore/cladestore.html`) and Deep Green at the Green Plant Phylogeny Research Coordination Group (`http://ucjeps.berkeley.edu/bryolab/GPphylo/`). There are separate databases that address phylogenetic reconstructions, such as TreeBASE (`http://www.treebase.org/treebase/`) . Finally, the Tree of Life (`http:// tolweb.org/tree/phylogeny.html`) project allows the user to explore the commonly agreed upon tree of life and retrieve information about particular taxa. However, each of these databases does not allow the user to explore data or do comparative studies about this data easily. Most phylogenetic data repositories pick one aspect of this complex data set, concentrating on storing information concerning this aspect. Besides creating an awkward situation for the user, where he or she must scour multiple repositories for the information he or she is looking for, it also causes some other signi cant problems.

This paper discusses the importance of phylogenetic data integration. Data integration can not only help with this problem of disconnection of phylogenetic information, but also help in developing more consistent tools for the eld. It examines brie y some data integration methodologies common in computer science. Finally, it suggests using exible data quality frameworks, such as BIO-AJAX, for attempting to begin the difficult task of creating environments to integrate phylogenetic data as well as develop interfaces for comparative studies for such data.

2 Phylogenetic Data

Phylogeny is a fast-growing eld with multiple applications both inside of the biology domain as well as outside of it. The primary purpose of phylogenetic information is to help accumulate data about the origin and nature of a taxon as well as allow the user to compare this information with similar taxa. Phylogenetic data represents evolutionary data with respect to some organism. This evolutionary data can be related to the Tree of Life or be a population study of the taxa. For example, with viral studies, phylogenies are used extensively to model contagion patterns [12].

Data concerning phylogenies incorporates many di erent structures. Primarily, there is text-based data. This can include both structured and unstructured data. Text-based information can incorporate anecdotal and experimental information about the taxon. For example, the Tree of Life has many pages that describe the behaviors and characteristics of

*Department of Computer Science, Montclair State University, Montclair, NJ, 07043, herbertk@mail.montclair.edu

†Department of Computer Science, NJIT, NJ, 07102, wangj@njit.edu

various taxa that experts have contributed for the repository. Besides the data about the taxon, there is usually annotation data in the form of study citations relating to the taxa [12]. Figures 1 and 2 demonstrate some of the complexity surrounding this data set. These gures do not entirely de ne the data of interest regarding phylogeny, but highlight the most common data concerns. Also, they give a glimpse of the complexity of the data. They reveal some of the cases where seamless integration can play an integral part in phylogeny research.

Lineage Paths for Homo sapiens

NCBI

root −> cellular organisms −> Eukaryota −>
Fungi/Metazoa group −> Metazoa −> Eumetazoa −>
Bilateria −> Coelomata −> Deuterostomia −> Chordata −>
Craniata −> Vertebrata −> Gnathostomata −>
Teleostomi −> Euteleostomi −> Sarcopterygii −>
Tetrapoda −> Amniota −> Mammalia −>
Theria −> Eutheria −> Primates −> Catarrhini −>
Hominidae −> Homo/Pan/Gorilla group −> Homo

ITIS

Animalia −> Chordata −> Vertebrata −>
Mammilia −> Theria −> Eutheria −>
Primates −> Hominidae −> Homo

Figure 1. Lineage path comparison between NCBI and ITIS.

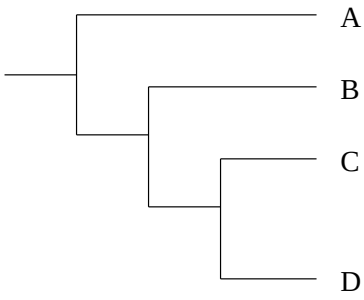

Figure 2. A Standard representation of a phylogenetic tree

Phylogenetic data usually includes information about comparative studies concerning the taxon or allows the user to access information. There are a number of types of information that can be included in this category. If this is an evolutionary study, then the species has a scienti c name based on the Linnaean Nomenclature. This scienti c name is representative of a taxons place in the Tree of Life. However, since the Linnaean Nomenclature is hierarchical, with the name representing evolutionary relationships that for some species are currently under debate, there are a number of projects looking to replace the Linnaean Nomenclature, such as PhyloCode [14]. Moreover, in every vernacular, species and taxa are usually referred to by their common name as opposed to the scienti c name. Therefore, there is a lot of information stored just within the taxon's name which also needs to be integrated across a number of vocabularies [12].

Figure 1 represents the lineage paths of a speci c taxon T, in this case T being *Homo sapiens*. The lineage path for T represents a model from which a repository or an organization recognizes as the path of nodes which are ancestors of T in a phylogenetic tree. Often this path represents T's placement in the Tree of Life. Figures 1 and 2 represent lineage path interpretations from both NCBI Taxonomy Database and the Integrated Taxonomic Information Resource for *Homo sapiens*. The taxon *Homo sapiens* represents the most research species and yet, as seen between the example lineage paths from the two databases, there are marked di erences. In the case of *Homo sapiens*, the di erent interpretations of the lineage paths represent how each database stores their data. For less researched species, there can be even greater di erences between the lineage paths [12].

Next, there is data about the composition of the taxon itself. This can includes information about the observable characteristics of a taxon or genomic and proteomic sequence data. Finally, there are the phylogenetic trees which characterize relationships between sets of taxa according to a phylogenetic reconstruction algorithm. Each of these also includes annotation data about the authors of the study as well as methodologies used for generating such data [12].

Figure 2 represents a generic phylogenetic tree R obtained from a phylogenetic reconstruction. Looking at R, it is obvious there are a number of layers of information within this tree structure that are of interest. First, the structure itself is of interest since the tree models the phylogenetic relationship between taxa A, B, C and D. Moreover, each taxon represents a point of interest. For each taxon, information concerning the nature of the taxon, genomic and proteomic data, peer reviewed studies and other phylogenetic trees it is a member of are quite possibly available and of de nite interest to

the user. However, most repositories can usually only support linking data within its own repository. A couple of resources, such as NCBI Taxonomy Browser attempt to integrate phylogenetic information from other repositories. However, this methodology, referred to as "Link Out databases, is limited by semantic problems. These difficulties arise from various abnormalities including database reliability and the taxonomy nomenclature [12].

3 Methodologies for Integration

Data integration is a well researched problem in computer science [1, 2, 4, 6, 8, 9, 10, 11]. Data integration is used for many purposes. First and foremost, data integration allows for the melding of multiple sources of data to give the user a seamless view of a set of data from multiple databases. From this functionality, several di erent uses can be inferred. For example, data integration can also help with preserving data quality. Since schema matching becomes an integral part of integration, algorithms for integrating the data can be used in the frame of exploratory data mining to detect abnormalities, inconsistencies and outliers within a data set spanning multiple repositories.

Within data integration research, there are many methodologies that can be used to perform the data integration. Moreover these methodologies have been classi ed in numerous ways. The pervasive methods for currently integrating biological data include data warehousing methods, mediator based methods and navigational methods [9].

Data warehousing methods are one of the most pervasive methods for integrating data. In this method, data from multiple sources is stored in one data warehouse. The schemas are matched during the data warehouse creation phase and data is stored based on this schema matching. All queries on the system are then enacted on the local warehouse. While this technique provides quick access to the data, it does have some drawbacks. Primarily, as the data corpus becomes larger, storage becomes an issue. Also, in the maintenance phases, if a source changes its schema, difficulties can occur in updating the warehouse [9].

Mediator based methods use query based approaches on the data sources to integrate the data. The data remains in the sources, but the system is able to extract the proper data through these views of the source data. Three pervasive methods exist of mediator based approaches. They are local as view (LAV), global as view (GAV) and a combination of the two methods called GLAV [1, 6, 11]. Mediator based methods o er exibility in that storage concerns are no longer a problem. However, depending upon the view of the data, query complexities as well as the nature of the local data model become key concerns. Currently, there is extensive research in this area to further match schema in a more meaningful way, as opposed to just a structural way. Semantic integration consists of extensive analyzing the source schemas and integrating them based on this analysis. This type of integration can use ontologies and helps to make the data ready for Semantic Web applications [4, 9].

Navigational techniques exploit the nature of on-line biological databases. Currently, most peer-reviewed biological data studies must have that data submitted to a freely accessible World Wide Web data repository. Most of the popular on-line biological data repositories also provide query mechanisms for users. Navigational techniques exploit this property of the repositories by using the query information the user submits to their database as a query for an external database. The connection is then represented through a hyperlink. This method of integration ensures data is current. Moreover, since the nature of hyperlinks is to create a connected graph between data sources, organizing and visualizing the data becomes each sources responsibility. However, navigational techniques require that for each data source, queries are speci ed properly and that each databases query mechanisms be maintained. If two sources use di erent nomenclatures, then the integration can fail. Moreover, since the navigational technique keeps the look at feel of each source, the seamless view of the data that more traditional integration techniques try to provide can be eliminated [9].

Each of these three techniques is currently used in a couple of phylogenetic repositories and general biological resources. The Tree of Life repository can be viewed as using the data warehousing technique, where data is collected through multiple sources and provided to the user. Moreover, the Tree of Life also uses a variation on the navigational technique to facilitate browsing the archive. However, since the data is stored locally, this navigational technique represents only the data within the Tree of Life data warehouse. Mediator based techniques can be seen in BioKleisli [3], a federated database for providing read access to multiple data sources with complex structured data critical to the Human Genome Project. Navigational techniques for integration are currently the most popular form of integration for phylogenetic data repositories. The most successful is the Link Out method at NCBIs Taxonomy Database. Link Out allows repositories outside of NCBI to enter queries to be performed on their repository. This query is provided as a hyperlink on the taxon information page at the Taxonomy

Database's Website. While this creates an excellent tool for users to view other data repositories information, it does su er from broken links as well as problems with query speci cations.

4 Phylogenetic Data Integration

Integrating phylogenetic data is not a trivial problem. Issues from how to match the data to how to represent the data arise within the area. When considering phylogenetic data, many questions arise concerning how to organize the integration of the data and which views of the data should be given.

Some issues to consider include:

- How should schema matching be handled, especially when working with complex data?
- How should the data be represented to the user so that it is communicated meaningfully and e ectively?
- What methods of integration are best for this data set?
- What query language and optimization techniques are suitable for accessing the complex data?

First and foremost concerning the data integration is how to match the data sources. This becomes extremely difficult due to the varying nature of the data as well as the complexity of the data. Some data sources only concentrate on speci c sets of taxa, while others are general repositories. Some repositories only work with text-based structured data, while others primarily work with phylogenetic tree data. Therefore, picking integration techniques become a delicate process.

To handle these integration issues, data quality research may be able to help with this problem. Ideally, due to the complexity of the data set, multiple integration techniques will ultimately be needed to address these problems. Therefore, another issue that is placed upon the developed is how to combine the integration techniques. Data quality frameworks can help with this issue. Data quality frameworks give a conceptual basis for which data quality techniques should interact with each other. Two possible frameworks that can aid in the integration are the popular Extraction, Translation and Loading Framework (ETL) and the BIO-AJAX framework.

The ETL framework [13] is the most popular framework for integrating data as well as evaluating and preserving data quality. The ETL framework consists of three phases for which it is named. In the extraction phase, the data of interest from the sources is acquired from the sources. In the translation phase, the data is modeled ino whatever form needed. Finally, in the loading phase, the transformed information in loaded into the database, ready for interaction with the user.

BIO-AJAX [7], a framework developed for managing biological data quality and based on the declarative framework created by Galahardas et. al. [5], has some similar properties to ETL. However, there are six clearly de ned conceptual operators from which the integration can be based upon. These operations include CLASSIFY, CLUSTER, MAP, MATCH, MERGE and VIEW. Each of these operations can be instantiate by the user with whatever algorithm the developer chooses, as many times as needed.

The bene t of both of these methods is that, ultimately, the developer can specify desired algorithms he or she needs to perform the integration. However, since these are conceptual frameworks, the algorithms work together in a speci c manner.

5 Conclusion and Future Work

This paper discusses data integration issues within phylogeny and mentions brie y integration techniques that are of interest for addressing this problem. For phylogenetic research to further advance, and for the eld to expand to include users that are not experts, serious integration strides need to be made. Moreover, with multiple elds, such as medicine, homeland security and computer security looking at exploiting these natural hierarchical models, integration becomes necessity for using this information.

Data quality methodologies can o er phylogenetic integration a possible method for aiding in integration projects. Two frameworks speci cally, the ETL framework and BIO-AJAX, o er exibility while giving the user a conceptual framework to develop the integrated tool. Recent progress on these concepts has included implementing BIO-AJAX to handle nomenclature integration into TreeBASE. From this project, nomenclature speci cation has become more exible within TreeBASE. Future projects include creating comparative environments for nomenclature study and further integrating currently available phylogenetic resources.

References

[1] A. Calí, G. De Giacomo and M. Lenzerini. Models for information integration: Turning local-as-view into global-as-view. In *Proc. of Int. Workshop on Foundations of Models for Information Integration*, 2001.

[2] S. Davidson, C. Overton and P. Buneman. Challenges in Integrating Biological Data

Sources, *Journal of Computational Biology*, 2: 557-572, 1995.

[3] S.B. Davidson, G. C. Overton, V. Tannen and L. Wong. BioKleisli: A Digital Library for Biomedical Researchers, *Int. J. on Digital Libraries*, 1(1): 36-53, 1997.

[4] A. Doan, N. Noy, A.Y. Halevy. Introduction to the special issue on semantic integration, *ACM SIGMOD Record*, 33(4):11-13, December 2004.

[5] 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.

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

[7] 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.

[8] K.G. Herbert, J. Westbrook and J.T.L. Wang. Data Integration in Biological Database, in *Proc. of the Atlantic Symposium on Computational Biology and Genome Information Systems & Technology*, Durham, North Carolina, September, 2003.

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

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

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

[12] R.D.M. Page and E.C. Holmes. *Molecular Evolution: A phylogenetic approach.* Oxford: Blackwell Scienti c, 1998.

[13] 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.

[14] C. Soares. Whats in a name? Scienti c American, November 2004.