A Structure-Based Search Engine for Phylogenetic Databases*

Huiyuan Shan[†]
Katherine G. Herbert[‡]
William H. Piel[§]
Dennis Shasha[¶]
Jason T. L. Wang[‖]

Abstract

Phylogenetic trees are essential for understanding the relationships among organisms or taxa. Many of the current techniques for searching phylogenetic repositories allow the user to perform a keyword-type search or an aligned sequence data search, or to browse a hierarchical list of taxa. Here we describe a new search engine that allows the user to present an example phylogeny, or a query tree, and then searches a phylogenetic database for trees that (approximately) contain the query structure. The presented search engine is fully operational and is available on the World Wide Web.

Keywords: Bioinformatics, phylogenetic trees, query processing, structure search, system.

*Work supported in part by U.S. NSF grants IIS-9988345 and IIS-9988636.

[†]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.

[§]Institute of Evolutionary and Ecological Sciences, Kaiserstraat 63, Leiden University 2311 GP Leiden, Netherlands. Current address: Dept. Biol. Sci., 608 Cooke, University at Buffalo, Buffalo, NY 14260, USA.

[¶]Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, USA.

[‖]To whom correspondence should be addressed: College of Computing Sciences, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA. Tel: (973) 596-3396, Fax: (973) 596-5777, Email: wangj@oak.njit.edu.

1 Introduction

Phylogenetic trees, usually represented by a dendrogram, model the evolutionary history of a set of taxa that have a common ancestor. The internal nodes within a particular tree represent older organisms from which their child nodes descend. The children represent divergences in the genetic composition in the parent organism. Since these divergences cause new organisms to evolve, these organisms are shown as children of the previous organism.

Currently, in studying the phylogenetic data, most systems use search methods that do not exploit the structure of the data. These systems usually adopt a keyword-based search tool, a sequence alignment, or a manually operated browser [1, 2, 3, 5]. The keyword-based tool allows the user to enter the name of a taxon or some identifying quality such as an identification number to search the database. It then returns all the data it has stored on that particular taxon. Some systems even allow the user to search more than one taxon at a time, using the traditional "AND" and "OR" operators to decide what set of data to return to the user. In sequence alignment, the taxon uses the genetic code from the sequence to align it with homologous sequences. If the sequences are similar, most likely they share a common ancestor. A phylogenetic tree can then be inferred from the pattern of nested instances of shared, derived mutations. Once the taxa are selected that align with the query taxon, links to other information about the returned taxa can be provided. Finally, browsing a hierarchical list allows users to examine one taxon at a time, learning only about that taxon.

Many of the systems that employ the search methods described above include visualization techniques that allow the user to view an entire section of a phylogenetic tree, or the entire tree, as well as interact with it. These interactions may involve visual inspection of the tree as well as linking to other trees that contain a particular taxon, or viewing isomorphism trees so that relationships between the taxa can be better understood.

However, none of the existing systems provides the user with the ability to search a database for the structure of a phylogenetic tree or structures similar to a query structure (as far as we know). Since the structure of a phylogenetic tree models very important information about the taxa contained within the tree, structure search becomes a very helpful as well as important tool for researchers studying phylogeny. This could, for example, help a researcher who desires to identify alternative evolutionary hypotheses for a set of taxa and how their evolutionary histories differ.

We describe here a new search engine, called TreeSearch, that allows the user to query by the structure of a phylogenetic tree. This search engine is fully operational and has been integrated into the phylogenetic information system, TreeBASE, developed at Harvard, UC Davis, Leiden University, and the University at Buffalo. Section 2 reviews TreeBASE. Section 3 presents TreeSearch. Section 4 concludes the paper and describes some future work.

2 TreeBASE

TreeBASE [3, 5], accessible at `http://www.treebase.org`, is a relational database containing phylogenetic information from research papers submitted to the Web site. This site then allows users to search the database freely according to various keywords, and see visual representations of the trees. Moreover, it allows the user to gain access to information concerning a tree as well as use comparison tools to learn more about various taxa contained within the tree and their relationships with other taxa within the database.

The dataset TreeBASE maintains consists essentially of phylogenetic trees submitted to TreeBASE by the authors of the papers that present the trees. The site accepts for review any peer-reviewed and published paper that presents information on any type of phylogenetic trees. For the paper to be contained within the database, it must be submitted to the site. The paper then goes through a review process before the submitted data are officially put within the database.

The schema and relational tables in TreeBASE contain various types of data including the citations of the papers stored within the database, the abstracts from the papers, the information about the authors, the algorithm used to obtain the phylogenetic trees, the titles and types of the trees, the software used to perform the relevant analysis, the association of the trees and matrices with the study through which they were obtained, and the information about the taxa.

TreeBASE allows the user to search its database by various keywords, including taxa, author, citation, study accession and matrix accession. Search results contain the information about the study that an input keyword was found in. This information includes the publishing date, the author, the title of the study in which the keyword was found, and the periodical in which the study appeared. Also, accompanying the study are analyses of the data presented within the study. These analyses can include the matrix from which the phylogenetic trees are generated [4], a link to drawing the tree in a frame within the Web site, a link to download the tree so that the user can view it on his or her own viewer, and a link to "mark" a tree which allows the user to store the tree for quick retrieval later. TreeBASE is also equipped with various visualization tools for drawing and displaying trees, and allows the user to "tree surf" and download the matrix of a particular tree. Detailed descriptions of TreeBASE can be found in [3].

3 TreeSearch

The source code of TreeSearch, accessible at `http://cs.nyu.edu/cs/faculty/shasha/papers/treesearch.html`, is developed independently for experimenting with structure-based searches in tree databases. The underlying algorithms and their applications to querying XML data can be found in [6]. We focus here on applying the code to phylogenetic database search. This structure-based search engine is now fully operational, accessible at `http://aria.njit.edu/~biotool`, and has been integrated into TreeBASE. To access it from TreeBASE, please visit the Web site `http://www.treebase.org/treebase/console.html`, and click "Structure" in the pull down menu (search window) on the Web site. TreeSearch is being visited a few hundred times a month.

3.1 System Architecture

Figure 1 shows the software architecture of TreeSearch. The system is composed of four components: Web-based Interface, Query Processor, Structure Viewer and Performance Log. From the Web-based Interface, the user is able to type in his/her own query (an exam-

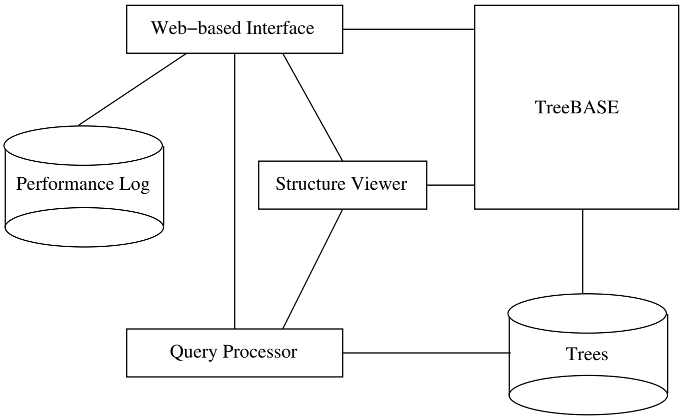

Figure 1: The software architecture of TreeSearch.

ple tree), upload the query tree from a file, or use and modify a sample query provided by the system. The Query Processor searches TreeBASE for phylogenetic trees (approximately) containing the query tree. The Structure Viewer displays the trees using either a parenthesized string notation or a dendrogram format, which are presented to the user via Web-based Interface. User queries and their timestamps are maintained in the Performance Log, which helps to analyze user's needs and better tune the system for working more effectively. TreeSearch is connected to TreeBASE on the Web and therefore it uses the visualization tools available in TreeBASE for displaying trees graphically. The system is implemented using Java, HTML, Perl CGI, and K (`http://www.kx.com`). The system can run under both Windows and UNIX.

3.2 Querying the System

Figure 2 shows the main screen from which the user can query the system. The uppermost frame displays the main menu for the system. The leftmost bottom frame allows the user to input a query. The rightmost bottom frame allows the user to gain access to TreeBASE; when a search is performed, this frame also displays the search results.

Figure 3 illustrates an example query and search results. The left window is the main screen, containing menu items, a query tree and search results. The two windows to the right of the main screen display the trees graphically; the top window shows the query tree and the bottom one shows a matching tree. To view a tree in a dendrogram format, the

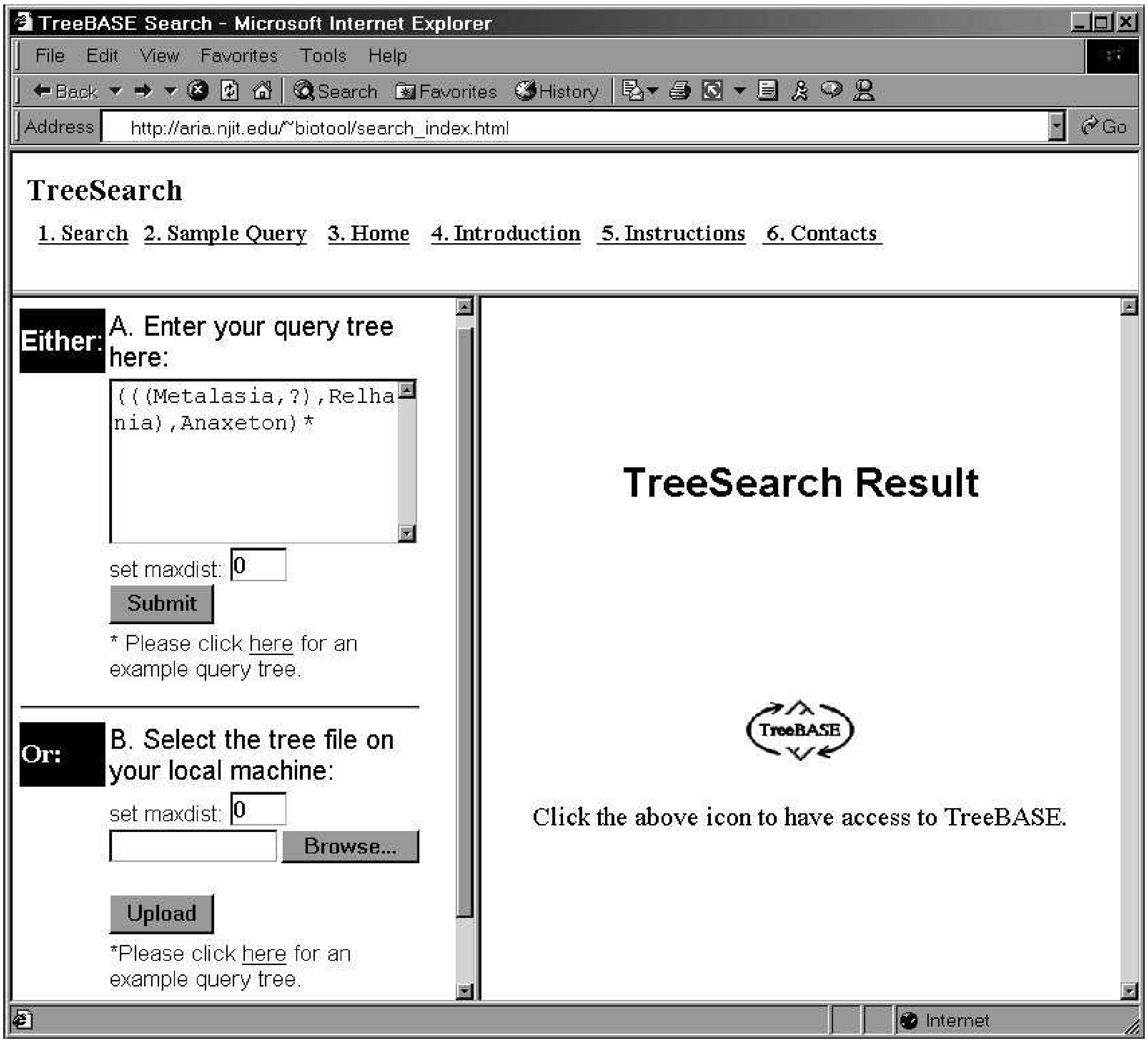

Figure 2: The main screen of TreeSearch.

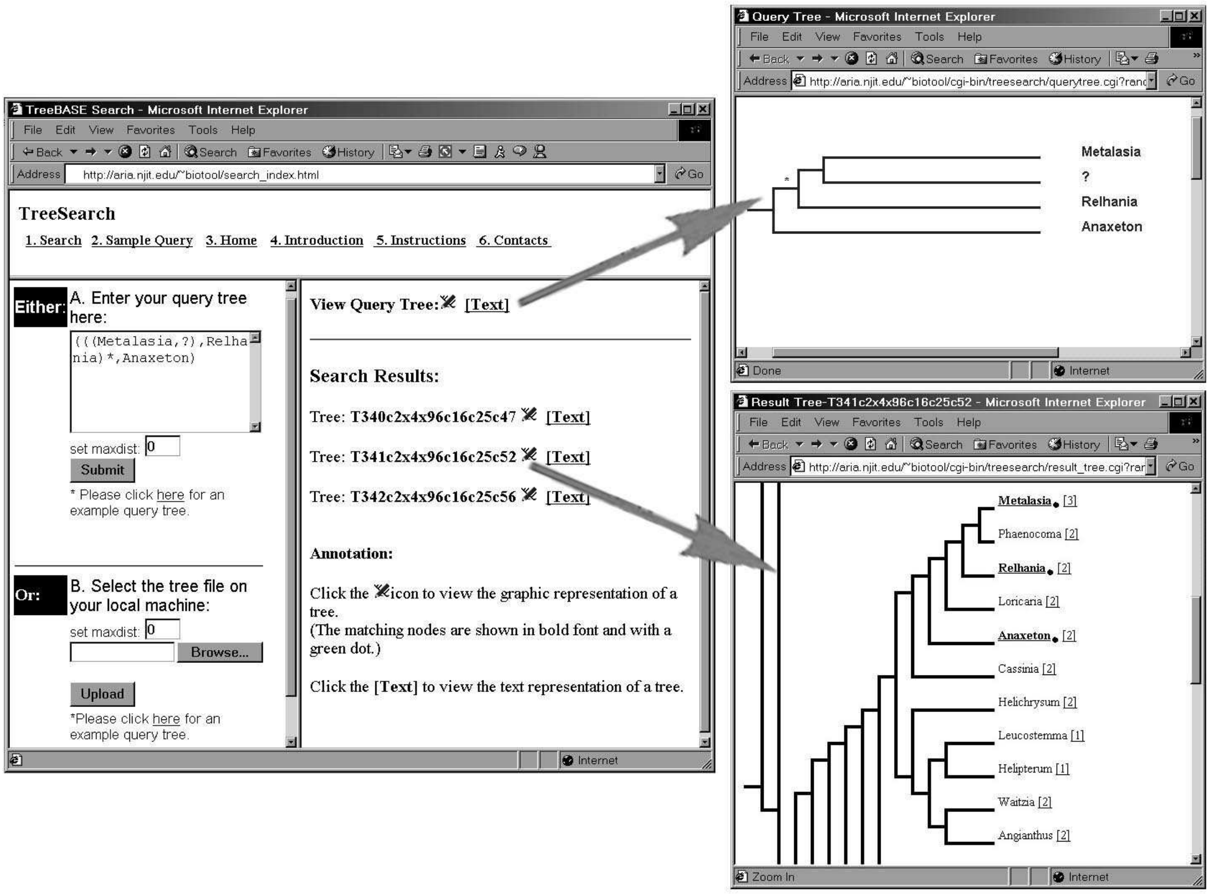

Figure 3: The TreeSearch interface for showing an example query and search results.

user would need to click the icon with the pencil overlaid upon the phylogenetic tree in the main screen. To view the parenthesized string notation, the user would need to click on the "Text" link in the main screen.

The query tree in Figure 3 contains a variable length don't care (VLDC), denoted "$*$", and a fixed length don't care (FLDC), denoted "?". When matching with a data tree, the VLDC "$*$" in the query tree may substitute for a path of length zero or more in the data tree.[1] The FLDC "?" in the query tree may substitute for a single node in the data tree.

In addition, the system allows the user to specify a distance value for performing approximate searches. Given an integer d (typed in the "set maxdist" window in the main screen), the search engine finds all the phylogenetic trees T in TreeBASE that approxi-

[1]The VLDC may also match with several paths, connected together with a shape like an umbrella of nodes in the data tree [7, 8].

mately contain the query tree Q within distance d. That is, T contains a substructure T' and the distance from Q to T' is at most d. We measure the distance from Q to T' by the total number of root-to-leaf paths in Q that do not appear in T'; the nodes in T' that do not appear in Q can be freely removed.

Referring to Figure 3, in the matching tree dendrogram, the entire phylogenetic tree is drawn and the user can use scroll bars to view portions of the tree. The specific taxa specified within the query tree are highlighted in the matching tree using underscored red font with a green circle next to the taxon's name. Also, each taxon has a number next to its name. This number represents the number of studies in TreeBASE the taxon is found within. If the user clicks on this number, he or she is linked to TreeBASE so that he or she can search on that taxon about those studies.

4 Conclusion and Future Work

TreeSearch provides the user with a powerful tool to allow the user to query phylogenetic information. By allowing the user to enter a structure as a query, he or she can obtain information about relationships between taxa that would otherwise require visual analysis.

The TreeSearch project is ongoing. Future work includes

- developing better user interface and visualization tools to help the user enter the query tree more easily and better understand the trees returned from the system;
- developing more effective schemes that mix and combine TreeSearch with the different search methods available in TreeBASE as well as other phylogenetic information systems;
- understanding the kinds of analysis that scientists do and extending TreeSearch in that direction; and
- studying various techniques for ranking search results for approximate searches and exploring the possibilities of using XML and metadata concerning phylogenetic information for improving searches.

References

[1] B. L. Cohen, J. A. Sheps, and M. Wilkinson. Archiving molecular phylogenetic alignments as NEXUS files. *Systematic Biology*, 47:495–496, 1998.

[2] D. R. Maddison. Tree of life. `http://phylogeny.arizona.edu/tree/phylogeny.html`.

[3] 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*, Tsukuba, Japan, 2000.

[4] M. A. Ragan. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, 1:53–58, 1992.

[5] 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, 1994.

[6] D. Shasha, J. T. L. Wang, H. Shan, and K. Zhang. ATreeGrep: Approximate searching in unordered trees. Submitted, 2002.

[7] J. T. L. Wang, K. Jeong, K. Zhang, and D. Shasha. A system for approximate tree matching. *IEEE Transactions on Knowledge and Data Engineering*, 6(4):559–571, 1994.

[8] K. Zhang, D. Shasha, and J. T. L. Wang. Approximate tree matching in the presence of variable length don't cares. *Journal of Algorithms*, 16(1):33–66, 1994.