

PROCEEDINGS PAPER

Darwintree: A Molecular Data Analysis and Application Environment for Phylogenetic Study

Zhen Meng¹, Hui Dong², Jianhui Li¹, Zhiduan Chen³, Yuanchun Zhou¹, Xuezi Wang¹ and Shouzhou Zhang²

¹ Scientific Data Center, Computer Network Information Center, Chinese Academy of Sciences, 4, 4th South Street, Zhongguancun, Haidian District, 100190 Beijing, China

zhenm99@cnic.cn

lijh@cnic.cn

zyc@cnic.cn

wxz@cnic.cn

² Key Laboratory of Southern Subtropical Plant Diversity, Fairylake Botanical Garden, Shenzhen & Chinese Academy of Sciences, 160 Xianhu Rd., Luohu District, 518004 Shenzhen, China

faye.huidong@gmail.com

shouzhouz@126.com

³ State Key Laboratory of Systematic and Evolutionary botany, Institute of Botany, Chinese Academy of Sciences, 20 Nanxincun, Xiangshan, 100093 Beijing, China

zhiduan@ibcas.ac.cn

DarwinTree (<http://www.darwintree.cn>) provides an integrated bioinformatics platform that supports all phases of the analytical pathway for phylogenetic study from data collections, phylogenetic tree constructions, visualization of the tree of life, and web-based rendering to specific application services & data mining. First, it is a repository for sequence records that form the basic data unit of all phylogenetic studies. Second, it is a workbench that aids the management, quality assurance, and analysis of phylogenetic data. Third, it provides a community of phylogenetic application services with tree reconstruction and related data mining. This paper provides a brief introduction to the key elements of DarwinTree, discusses their functional capabilities, details the database features, phylogeny pipeline, data mining tools, and specific application services available to support it, and concludes with future prospects.

Keywords: Phylogenetic study; Data collection; Tree reconstruction; Data mining tools; Specific application services; Taxonomy; DarwinTree

1 Introduction

The work of phylogenetics includes studies at many levels that make use of databases in varying degrees, from simple searches, to checks for PCR contamination, to finding homologs for a given sequence among outgroups, to more comprehensive studies based on data mining large numbers of taxa or loci (Bininda-Emonds et al., 2008; Francesca D. Ciccarelli et al., 2006; C. H. Li, Orti, Zhang, & Lu, 2007; McMahon & Sanderson, 2006; Sanderson, Boss, Chen, Cranston, & Wehe, 2008). The phylogenetic tree is a tree-like branching diagram that represents various (classes) kinships among organisms through the study of biological sequences to infer the evolutionary history of the species in phylogenetics. The tree can be constructed from a sequence alignment including DNA sequences and protein sequences or a protein structure comparison including rigid structure and multi-laminated structure. The studies based on molecular evolution have been applied to many areas, such as gene evolution, fauna division, mating systems, species identification, paternity testing, environmental monitoring, and the disease sources shifted among species. Meanwhile, the Tree of Life (TOL) links the huge amount of information known about all living species (existing and extinct) in a phylogenetic tree. It can be used to clarify the origin of life, biological evolution patterns, categories of biological

evolution and phylogenetic relationships, and the survival of biodiversity patterns and dynamics of change. Constructing a TOL and fully exploiting its information is another challenge faced by life sciences (Ciccarelli et al., 2006; Meng, Lin et al., 2011; Wapinski, Pfeffer, Friedman, & Regev, 2007).

One major achievement over the past 20 years in the field of phylogenetics (in particular TOL), concerns molecular data (especially, gene sequence data). As an example, release 159 from GenBank (a member of the International Nucleotide Sequence Database Collaboration (INSDC)) contains 75 billion nucleotides in 72 million sequences plus another 93 billion nucleotides just in the WGS (whole genome shotgun) sequence division stemming from 787 registered genome projects (Benson et al., 1999; Benson, Karsch-Mizrachi, Lipman, Ostell, & Sayers, 2011; Rindone, 1983). Another achievement is the advances made in calculation algorithms and computing infrastructure. Many tools and algorithms have been used in phylogenetics to further enhance the capabilities of large data processing. However, there are still many technical problems to be solved, including data collection and screening of DNA data, automatic construction of large trees (Supertree), further excavation and sharing of information, and so on (Ciccarelli et al., 2006; Li, Meng, Hou, Zhou, & Gao, 2013; Meng, Li, et al., 2011). The following are two different ways to approach these problems: 1) integrate a number of trees into a synthetic large tree, based on the overlap of two or more smaller trees or 2) construct the TOL directly on the super-data matrix for analysis. No matter which way is chosen, similar problems must be faced: 1) how to make full use of the existing public DNA sequence information databases; 2) how to effectively filter the information; 3) how to quickly auto-generate a tree of life that reflects the evolutionary history of different biological taxa; 4) how to fully exploit and utilize the tree of life implied in the huge amount of information.

DarwinTree (<http://www.darwintree.cn>), which was preceded by the Phylogenetic Analysis of Land Plants Platform (PALPP) (Meng, Lin, et al., 2011), was initiated to construct a molecular data analysis and application environment for phylogenetic study. It was compiled collaboratively by 3 organizations from the Chinese Academy of Sciences: the Computer Network Information Center (CNIC), the Institute of Botany (IB), and the Shenzhen Fairy Lake Botanical Garden (SZBG). DarwinTree is expanding its development on a global scale to build an international research alliance. It has initiated its first international campaign with the University of Florida for the genera phylogeny of Angiosperms.

Currently, DarwinTree provides an integrated bioinformatics platform that supports all phases of the analytical pathway for phylogenetic study, beginning with data collection and including gene data acquisition and management; phylogenetic tree construction, including sequences alignment and data matrix construction; reconstruction of different branches; assembly of large trees and modes of optimization; visualization of the tree of life and web-based rendering; and specific application services and data mining. In the remainder of this article, we examine the key elements of DarwinTree and detail the database features, phylogeny pipeline, data mining tools, and the specific application services available to support it, before concluding with a brief discussion of future prospects.

2 Key Elements

DarwinTree was established to reconstruct the environment and is committed to providing solutions for the study of phylogenetics. First, it is a repository for sequence records that form the basic data unit of all phylogenetic studies. Second, it is a workbench that aids the management, quality assurance, and analysis of phylogenetic data. Third, it provides a community of phylogenetic application services with tree reconstruction and related data mining. The key elements of DarwinTree are described in the following figure.

In order to support all phases of the analytical pathway for phylogenetic study, the modules are organized as in **Figure 1**. First, the basic data are prepared synchronously with an analysis of public bioinformatics resource databases, such as GenBank, Gene Ontology (Gene Ontology Consortium, 2004), etc. Each sequence is organized with “sequence feature” as the storage unit. Biologists can then select species to obtain complementary sequences according to their peer research status. Second, data washing is done to remove the pseudogenes, horizontal transfer genes, organellar copies, and recent paralogs. Third, the workflow of phylogenetic tree construction including multiple sequence alignment and data matrix construction, different branches reconstruction, assembly of large trees, and modes of optimization is carried out. Next, the results are integrated with information about morphology and fossils and other data by using web-based rendering visualization of the tree. Finally, a specific species library information platform (for example, the platform of apomixis in ferns (Liu et al., 2012)) can be built and related data mining done.

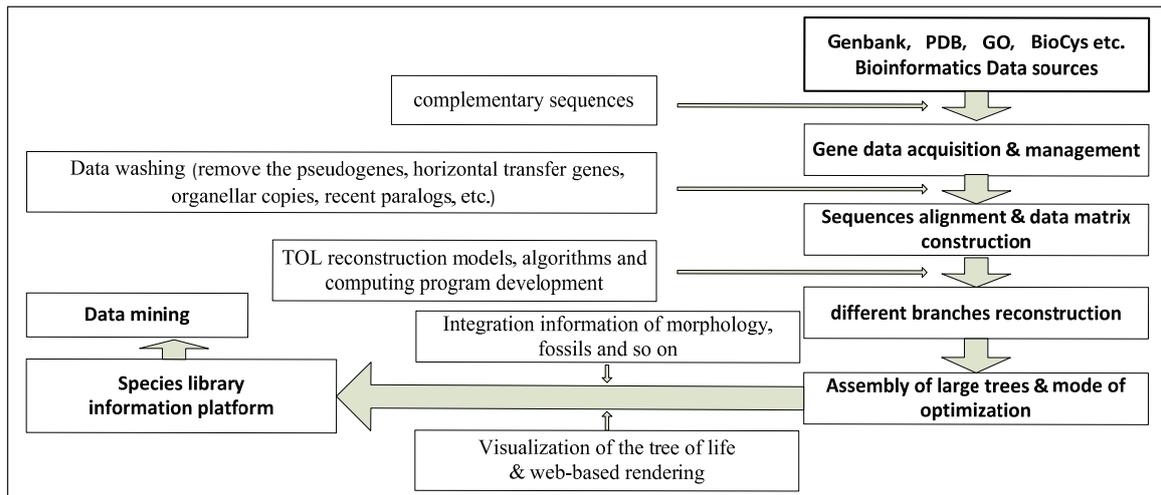


Figure 1: Modules of the DarwinTree: 1. synchronization with public bioinformatics data sources; 2. gene data acquisition and management; 3. sequences alignment and data matrix construction; 4. different branches reconstruction; 5. assembly of large trees and modes of optimization; 6. species library information platform; 7. data mining.

2.1 Database Features

All the sequence data in DarwinTree, including the primary data extracted synchronously with the public bioinformatics resource databases and the complementary sequences from biologists themselves, can be browsed in different levels of the taxonomic hierarchy through the taxonomy trees. There are two kinds of data models used in DarwinTree: 1) adding to one's own data in ongoing projects while retaining the quality of the public sources and one's own data and 2) building an effective private taxonomy system in "My DarwinTree" (Gao et al., 2011). In order to enhance the organization of the basal sequence data to be efficient and definable, a scalable, high performance, open source, document-oriented, key/value based database, MongoDB (<http://www.mongodb.org/>), is used to store data from public bioinformatics resource databases. It is updated every day.

2.1.1 Data Repository

DarwinTree supports the reposting of private sequence data and integrates them with public data. The server will evaluate the sequences and give evaluation reports on each sequence. After users submit sequences, the DarwinTree staff reviews the submitted information and returns the record by e-mail for users to review. Trace files and primers databases are also included in DarwinTree.

In the storage design of DarwinTree in MongoDB, the metadata (Meta) and sequence data (Sequence) are stored separately because of the weakened dependency of the NOSQL database. The fast query table (Query) is designed to meet the needs of real-time queries. A query goes first to the Query collection, which then retrieves the original data from the Meta and Sequence collections. The collections Meta, Sequence, Query, Statistics, and Taxonomy are associated through accession, but this association is not mandatory. To improve efficiency, each collection is indexed for Accession, Scientific Name, Taxon, and Gene in a manner that can be implemented in parallel queries in data partitioning storage (5 slices). To ensure data security, there are also multiple copies (3 copies) in storage.

Meanwhile, a "FEATURES", such as "gene", is separated out as a key/value pair, with its length and location information located at the sequence called "ACCESSION" in the GenBank. One sequence may be separated into several key/value pairs according to its "FEATURES". Because of the separation, the sequences are more appropriate for phylogenetics study, especially DNA barcoding study (Schindel & Miller, 2005), which uses sequences of specific marks. Therefore, the datasets from DarwinTree are superior to those from GenBank for use in phylogenetic study. Sequence data can also be obtained according to their length, and each sequence of a specific mark is appropriate for the location at the accession. For example, a sequence can be obtained as a gene mark "rbcl", with location "58339..59766" and length "59766-58339+1=1428" at DarwinTree. Compare this with the whole sequence being identified with accession "JN867587" at GenBank. For example, at the end of February 2014, there were up to 9,482,915 records of Gene Mark prepared for phylogenetic application.

2.1.2 Taxonomy Browsers

Currently, DarwinTree has two public taxonomy versions in its browsers, the NCBI Version and the APG Version. The NCBI version is based on and is synchronous with the NCBI taxonomy database. The APG version is compiled by the DarwinTree Phylogenetic Group. It is a revised genera list of Chinese land plants based on several references, such as APG III (<http://www.mobot.org/mobot/research/apweb/welcome.html>), Flora of China (<http://hua.huh.harvard.edu/china/>), ITIS (<http://www.itis.gov/>), and Wu et al.(2003). New public taxonomy versions will be added as the need arises. DarwinTree can also provide the “My Tree” function, which enables users to upload their own taxonomy trees to private accounts. In this situation, users need focus only on their own taxon.

2.1.3 Data Statistics and Customization and View

DarwinTree is designed for data analysis to accommodate large numbers of terminals, on the level of species, genera, order, family, tribe, or class. DarwinTree provides users with several tools for the management of mass sequence data, which include “Data Statistics”, “Data Customization”, “Data Subscription”, etc. The “Data Statistics” function allows users to do statistical analysis on the sequences storage status. Users can easily upload an .xls file with the name list. The servers will determine the number of sequences a user has deposited at deposit in the database and export that result to an .xls file. Visualization charts (flash) are also used for every group (taxon name). This is one of the unique features of DarwinTree and is quite useful to phylogenetic scientists before they sample the study. The “Data Customization” function allows users to customize their study group and specific a DNA Mark (gene). Users can conveniently view the latest customized data in their own workspace with the statistics charts of the customized data and the detailed information of each sequence. The sequences data can be viewed from the taxonomy tree and can be sent to the analysis platform in DarwinTree (see **Figure 2**).

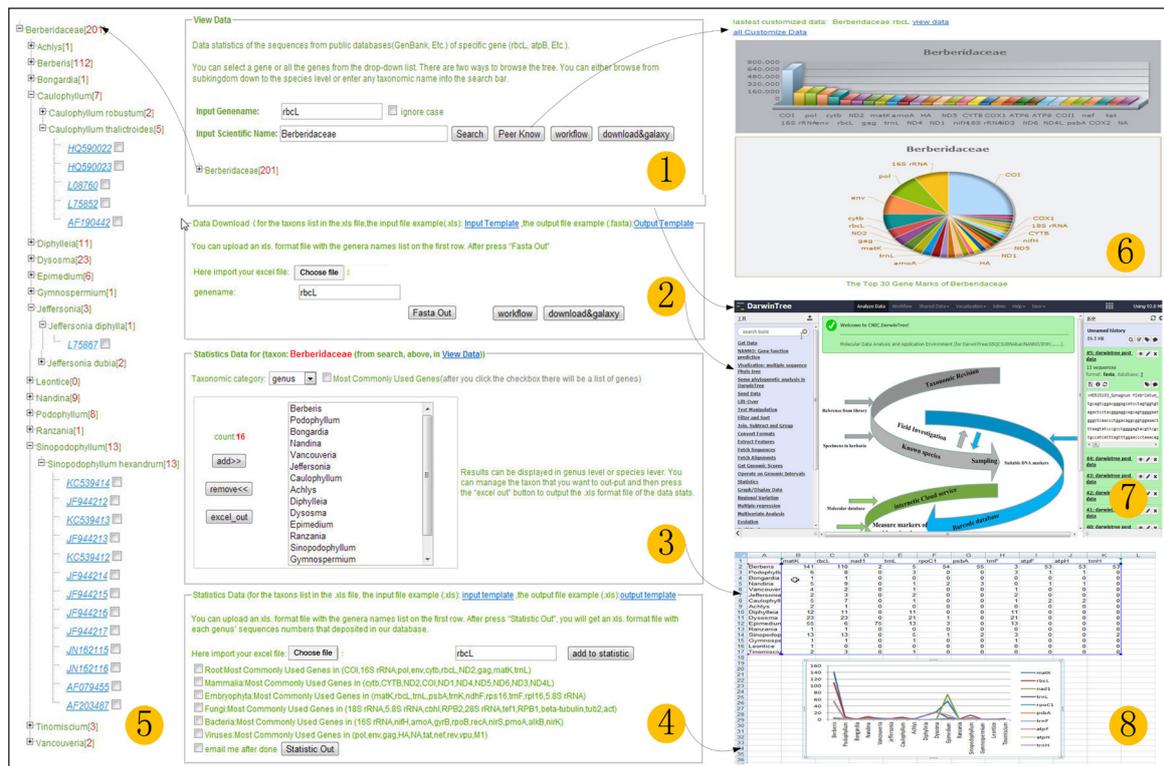


Figure 2: Data View in DarwinTree: 1) data search; 2) data download according to users' taxonomy with an .xls file input; 3) data statistics according to public taxonomy; 4) data statistics according to users taxonomy with input of an excel file and any sequence gene mark; 5) sequence data view according to taxonomy and any sequence gene mark; 6) “peer known” of taxon (any taxon of the top 30 gene mark sequences); 7) data submission to Galaxy or local workflow for analysis; 8) output of sequences data statistics with each mark.

2.2 Phylogeny Pipeline

Along with analysis in clusters with command line interface and some phylogenetic workflows integrated with Galaxy (Goecks, Nekrutenko, Taylor, & Galaxy Team, 2010), DarwinTree provides users with a web interface that allows them to do data batch extraction, data analysis with web-based bioinformatics tools, and phylogeny workflows management. It is accessible using several modern web browsers, such as IE, Firefox, and Opera. It is coded in Java and well designed in an MVC pattern using Spring Framework, one of the most popular lightweight J2EE frameworks (Lin et al., 2010).

Users begin the workflow with “Data Batch Extraction” or any other step from the “Toolkits Organization”. Currently, DarwinTree integrates several programs in the phylogenetic tree construction pipelines, including BLAST (Altschul et al., 1997; Meng, Li, et al., 2011) for sequence similarity search, ClustalW (Thompson, 1994) and its MPI implementation (Li, 2003) for multiple alignments, PhyML (Guindon et al., 2010), RAxML-VI-HPC (Stamatakis, 2006), FastTree 2 (Price, Dehal, & Arkin, 2010), MUSCLE (Edgar, 2004), and MrBayes 3.2 (Ronquist et al., 2012) for tree building and ATV (applet) (Zmasek & Eddy, 2001) for tree visualization.

Users click “Start Wizard Run”, which enables sequence analysis and editing during tree reconstruction. DarwinTree also provide a “One-click Run” button to initiate an automatic phylogeny analysis process. After choosing the multiple alignment and phylogeny tree reconstruction methods, users’ requests will run automatically in the server. All the software is set up with default options and parameter values, which can be modified in the “Customize Workflow” webpage. In addition, users can pause the process in the task list and check the results after the server has sent an inform email when the tree reconstruction process is finished. All the files generated during the process are downloadable (.fasta, .aln, .nex, .nex.con, .phy, .tree, .txt). Gene combination analysis is also supported in DarwinTree. Users can either upload several separate FASTA files or select different genes in our database for analysis.

2.3 Data Mining Tools and Specific Application Services

To promote the power of DarwinTree for phylogenetic study, several data mining tools and specific application services were developed and provided for the public. For example, the Gene Sequence Quality Control Tool(GSQCT) (Meng, Li, et al., 2012) and its parallel implementation Cloud-GSQCT (Meng, Xiao, et al., 2012) for data washing; the efficient MapReduce program for bioinformatics applications (bCloudBLAST) (Meng, Li, et al., 2011), PartFastTree, which constructs large phylogenetic trees and estimates their reliability (Li et al., 2013) and Rapidtree, which is the solution to rapid reconstruction of the phylogenetic tree (Meng et al., 2014) for improvement of bioinformatics tools; in terms of specific application tools, iDNABar, the rapid species identification toolbox for DNA barcoding collection, preservation, identification, and tracing (Meng, Zhou, Li, Gao, & Shen, 2012) and the website of information on apomixis (= apogamy) in ferns and lycophytes (Liu et al., 2012); and SoTree, Species Identification Service (SPI), the automated phylogeny assembly tool for ecologists etc. for data mining of big phylogenetic trees. These features are described in the following subsections.

2.3.1 GSQCT: A Gene Sequence Quality Control Tool

GSQCT is promoted as a solution to screening gene sequence data. It first extracts the initial datasets using gene annotation information, and then it calculates the content of uncertain characters in the gene sequencing for sequencing quality detection, to detect stop codons, to avoid pseudogenes, to detect custom serial strings, to remove contaminative sequence fragments, to do protein similarity calculation with the template protein of the object gene for homology detection, and finally to decide whether to select using a pre-determined threshold range, gene by gene. The screening result report is given (shown in **Figure 3**), and a multiple sequence alignment can be done to verify the homology with those verified sequences. This solution overcomes the problems of error or ambiguous annotations and uneven sequencing accuracy in the existing gene data filtering, which could lead to incorrectly constructed phylogenetics trees. The solution evaluation is introduced and shows good accuracy and effectiveness. Parallel implementations with Hadoop (Map / Reduce) for download can be found at: <http://www.darwintree.cn/common/tools.shtml>. The web service entrance is: <http://phylo.csdb.cn/gsqcs/>.

2.3.2 bCloudBLAST: An Efficient Mapreduce Program for Bioinformatics Applications

The BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities. DarwinTree presents an improved MapReduce-parallel implementation by splitting both input query sequence files and sequence databases for a search, called bCloudBLAST, which illustrates very good

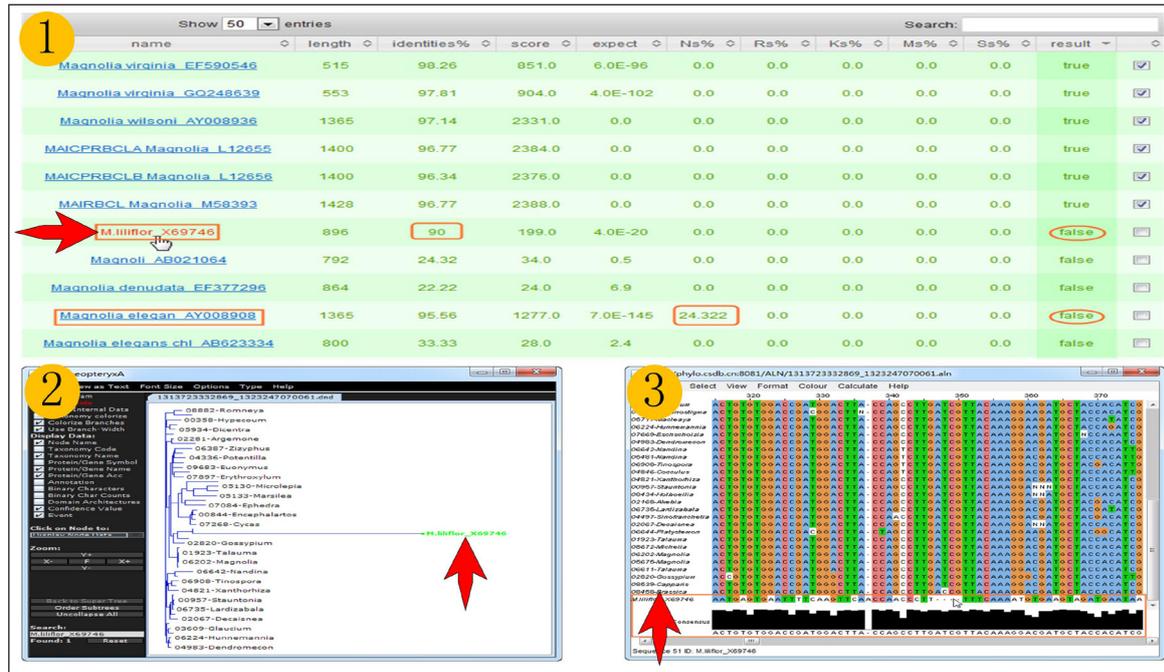


Figure 3: GSQCT Reports. The sequence needed for verification is marked “F” (accession: X69746, shown in 1). It is clear that it is the sequence marked “F”, whose genetic distance (divergence time) is much larger than any in the homologous gene sequence clusters (shown in 2) and whose key point is also not matched to the homologous gene sequence clusters’ alignment (shown in 3).

scaling and speedup behavior on large sequence databases. bCloudBLAST is written in Java, executable on UNIX/Linux, Windows, and MacOS systems. MapReduce and Hadoop libraries can be found at <http://hadoop.apache.org/>, and bCloudBLAST can be downloaded from <http://www.darwintree.cn/common/tools.shtml>.

Currently, bCloudBLAST using Hadoop has been adopted for use in DarwinTree with about 16 normal machines, each with eight Processors of 8*2.5G to screen data for phylogenetic analysis. However, some points need to be improved, including the method to split databases and parse the reports of BLAST. More evaluation will be done in the future to adapt to more gene and protein sequences with differing lengths.

2.3.3 iDNABar: A Rapid Species Identification Toolbox for DNA Barcoding

DNA barcoding technology, which uses one or a few DNA fragments for rapid identification of a species, is a research focus in phylogenetic biodiversity information studies. Therefore, iDNABar, a bioinformatics application toolbox, was designed to support all phases of the analytical flow and form a specimen collection to tightly validate the barcode library with any DNA fragment. It serves as a repository for specimens, analytical middle data, and sequence records that form the basic data unit of all barcode studies. It also serves as a workbench with management, quality assurance, and analysis of barcode data. It is a system with flexible security, web based delivery, and full interface with molecular biology databases, such as GenBank, BOLD (Ratnasingham & Hebert, 2007), ITIS (<http://www.itis.gov>), DarwinTree itself, and so on. It serves as a toolbox standalone for download in <http://www.darwintree.cn/common/tools.shtml>, which can run directly on local computers equipped with Apache (<http://www.apache.org>) and Mysql (<http://www.mysql.com>).

Using iDNABar, a researcher can manage data in all phases of the identification analytical flow based on DNA barcoding technology, from specimen collection to a tightly validated barcode library with any DNA fragment instead of a few limited Marks or DNA Barcodes. It promotes the methods of identification and data quality assurance. Key elements for management and sharing are also provided, including user security, statistical functions, and application interfaces. There are several applications based on iDNABar, such as the Professional Database for Freshwater Fish Species Identification in China (PDFFSIC, <http://www.fwfsi.csdb.cn>) and the Plant Barcode of Life (PBL, <http://pbl.csdb.cn:8080/>). With further expansion of the application, iDNABar will provide cloud services in which users need browsers having only scalable computing power, research summary update data in real-time alerts, and external visual display of personalized applications. Moreover, some methods for species identification based on multi-gene (multi-barcode) will be added.

2.3.4 RapidTree: A Solution to Rapid Reconstruction of the Phylogenetic Tree

Masses of data have accumulated with the development of DNA barcoding technology for the purpose of rapid identification using biological material samples. This is needed for construction of a big phylogenetic tree that includes numerous species, but it also raises a difficult question about how to find a location in the tree quickly using a sample sequence representing a specific species. RapidTree presents a solution based on an initial phylogenetic tree construction and a fast homology sequence alignment algorithm, which can quickly and automatically meet the needs of rapid species identification and big phylogenetic tree growth through a mark sequence, through a pipeline that includes organization of the basal dataset, construction of the underlying phylogenetic tree, construction of a basal similarity comparison library, reconstruction of the phylogenetic tree, and visualization of the phylogenetic tree in a web interactive environment. RapidTree software can be downloaded from the URL: <http://www.darwintree.cn/common/tools.shtml> (see **Figure 4**).

RapidTree presents a solution to rapid reconstruction of the phylogenetic tree and gives an implementation in the web environment that shows good scaling and speedup behavior in big tree construction and visualization with large sequence data. A NOSQL database, MongoDB, is used to improve the organization of the basal dataset's efficiency and definition. However, some points need to be improved in the future. The support and branch length will be evaluated and added in inserted node data. Other parallel methods of constructing the underlying phylogenetic tree are under study, waiting for evaluation and application.

2.3.5 SoTree: An Automated Phylogeny Assembly Tool for Ecologists

Ecologists have long recognized the importance of incorporating phylogenetic data in their work. Entire areas of study, such as community phylogenetics and comparative analysis, require detailed phylogenetic as well as ecological information (Cavender-Bares, Kozak, Fine, & Kembel, 2009; Webb, Ackerly, McPeck, & Donoghue, 2002). However, despite vast amounts of sequence data, progress in these fields has been slowed by the level of expertise required to create reliable phylogenies. Although there has been a recent explosion

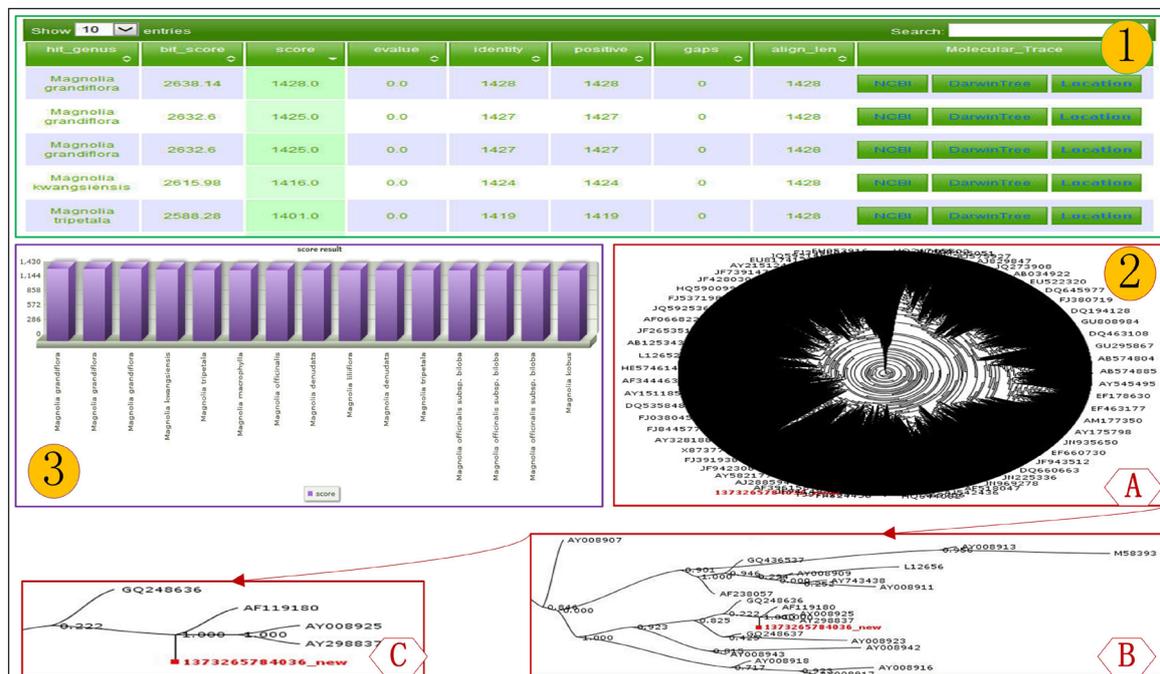


Figure 4: Reports possible using RapidTree. Typical reports from the use of the tree identification features of Rapid Tree are shown in 1. It lists the identification score, scientific name, positive score, and so on. Meanwhile, there are the links of the specific “Hit_genus/Hit_species” to COL (Catalogue of Life, <http://www.catalogueoflife.org>), and BHL (Biodiversity Heritage Library, <http://www.biodiversitylibrary.org>), the links of “Molecular trace” to sequence databases, NCBI and DarwinTree sequence details page, and the link of “Location” to the location at the big tree in the report web page. The rank of the identification score is presented as shown in 3. The visualization of the location at the big tree is shown in 2, in which the location becomes gradually more detailed moving from (A)→(B)→(C) and the location can be viewed via dragging of the mouse.

in the creation of extremely large phylogenies with many species (Izquierdo-Carrasco, Smith, & Stamatakis, 2011; Smith, 2009), there is often a mismatch between the species sequenced to build such trees and the species in which ecologists are interested (Pearse, Purvis, & Paradis, 2013).

SoTree (<http://www.darwintree.cn/flora/index.shtml>) is an online phylogenetic query tool where users submit a list of taxa (e.g., from an ecological community) with genus and species names, which returns a phylogenetic hypothesis for relationships among taxa. Any set of stored phylogenies, or a user-supplied one, can be chosen as the basis for the returned phylogeny. Several input models can be used, and output formats for the tree can be selected while any name will be marked at the accepted name or synonym according to the Flora of China (FOC). Currently, the source databases cover vascular plants. This offers a way to make phylogenies for non-specialists and provides an easy way for biologists to begin to move towards the evolutionary viewpoint.

3 Conclusion

DarwinTree consists of a repository for the sequence records that form the basic data unit of all phylogenetic studies and a workbench that aids the management, quality assurance, and analysis of phylogenetic data. It also provides a community of phylogenetic application services with tree reconstruction and related data mining. In this paper, the key elements of DarwinTree are presented, and the database features in MongoDB are described. The phylogeny pipeline, data mining tools, and specific application services available in DarwinTree are also introduced. These include the Gene Sequence Quality Control Tool (GSQCT) and its parallel implementation for data cleaning, the efficient MapReduce program for bioinformatics applications (bCloudBLAST) that improve the bioinformatics tools of BLAST, the solution to rapid reconstruction of phylogenetic tree (Rapidtree) for constructing large phylogenetic trees, the rapid species identification toolbox for DNA barcoding collection, preservation, identification, and tracing (iDNABar) for barcoding phylogeny and so on.

Along with the tree of vascular plants that has been recently completed in China, SoTree, an automated phylogeny assembly tool, can be used for ecologists. It allows anyone to generate a phylogeny by searching for species from an excellent genus-level reference phylogeny on the basis of taxonomy. Moreover, the sequences of the non-coding region mark have already been separated as a key/value pair, and these will be able to be viewed and used as gene mark sequences, according to their taxonomy (any mark any taxon) on the web in the near future. At the same time, the general objective of the tree of life (the largest phylogenetic tree) is to reconstruct the evolutionary history of all living species so that every biological organism can find its place in the big tree. Eventually the tree of life will include all kinds of levels of biological organisms, and perhaps then, the identification of biology organisms will be able to be conducted by barcode (the specific mark sequence). By that time, people will be able to identify every species ever named and discover new species according to one or several sequences using a handheld species analyzer. Although this is just a fantasy and there are still many problems with DarwinTree and all the other projects, there is a bright future for the theory and practical implementation of phylogenetic study.

4 Acknowledgments

All the authors would like to express their great thanks to the authors of many tools used by DarwinTree. We also thank Douglas Soltis and Pamela Soltis from the University of Florida for their generous help in many ways. Also all the work was supported by the Natural Science Foundation of China (NSFC) under Grant Nos. 91224006, 31270268, and 61003138, the Ministry of Science and Technology project 2014CB954100, the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDA06010202, the "Twelfth Five-Year" Plan for Science & Technology Support under Grant No. 2012BAK17B01-1, the Special Project of Informatization of the Chinese Academy of Sciences in the Twelfth Five-Year Plan under Grant No. XXH12504, XXH12503-05-01, the Five Top Priorities of "One-Three-Five" Strategic Planning, CNIC (No. CNIC_PY-1405), and the Shenzhen Science and Technology Innovation Council Funding (No. KQC201105310009A).

5 References

- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17), p 3389.
- Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F. F., Rapp, B. A., et al. (1999) GenBank. *Nucleic Acids Research* 27(1), pp 12–17.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2011) GenBank. *Nucleic Acids Research* 39, pp D32–D37.
- Bininda-Emonds, O. R. P., Cardillo, M., Jones, K. E., MacPhee, R. D. E., Beck, R. M. D., Grenyer, R., et al. (2008) The delayed rise of present-day mammals (vol 446, pg 507, 2007). *Nature* 456(7219), pp 274–274.
- Cavender-Bares, J., Kozak, K. H., Fine, P. V. A., & Kembel, S. W. (2009) The merging of community ecology and phylogenetic biology. *Ecology Letters* 12(7), pp 693–715.
- Ciccarelli, F. D., Doerks, T., Von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311(5765), p 1283.
- Edgar, R. C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1), p 113.
- Francesca D. Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J. Creevey, 1, et al. (2006) Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *SCIENCE* 311, p 1283.
- Gao, Y., Meng, Z., He, X., Liu, Y., Zhou, Y., & Li, J. (2011) A solution to integrate data for phylogenetic research. *iCBBE*(1), pp 1–4.
- Goecks, J., Nekrutenko, A., Taylor, J., & Galaxy Team, T. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11(8), p R86.
- Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* 59(3), pp 307–321.
- Izquierdo-Carrasco, F., Smith, S. A., & Stamatakis, A. (2011) Algorithms, data structures, and numerics for likelihood-based phylogenetic inference of huge trees. *BMC Bioinformatics* 12(1), p 470.
- Julie D. Thompson, D. G. H. a. T. J. G. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22(22), pp 4673–4680.
- Li, C. H., Orti, G., Zhang, G., & Lu, G. Q. (2007) A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. *Bmc Evolutionary Biology* 7.
- Li, J., Meng, Z., Hou, Y., Zhou, Y., & Gao, Y. (2013) PartFastTree: Constructing Large Phylogenetic Trees & Estimating Their Reliability. *2013 9th International Conference on Natural Computation*.
- Li, K. B. (2003) ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics* 19(12), pp 1585–1586.
- Lin, X., Meng, Z., He, X., Liu, Q., Liu, Y., Li, J., et al. (2010) A solution to integrate the phylogenetic tree's generation based on web. *iCBBE*(1), pp 1–3.
- Liu, H.-M., Dyer, R. J., Guo, Z.-Y., Meng, Z., Li, J.-H., & Schneider, H. (2012) The Evolutionary Dynamics of Apomixis in Ferns: A Case Study from Polystichoid Ferns. *Journal of Botany*, 2012, p 11.
- McMahon, M. M., & Sanderson, M. J. (2006) Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. *Systematic Biology* 55(5), pp 818–836.
- Meng, Z., Li, J., Zhou, Y., Cao, W., Xiao, X., Zhao, J., et al. (2012) GSQCT: A solution to screening gene sequences for phylogenetics analysis. *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery* 6, pp 2941–2945.
- Meng, Z., Li, J., Zhou, Y., Liu, Q., Liu, Y., & Cao, W. (2011) bCloudBLAST: an Efficient MapReduce Program for Bioinformatics Applications. *BMEI* 4, pp 2085–2089.
- Meng, Z., Lin, X., He, X., Gao, Y., Liu, H., Liu, Y., et al. (2011) Construction of the Platform for Phylogenetic Analysis. *Data Driven e-Science, ISBN 978-1-4419-8013-7. Springer Science+ Business Media, LLC*, pp 507–514.
- Meng, Z., Shao, J., Cao, W., Li, J., Zhou, Y., Wang, X., et al. (2014) RapidTree: a solution to rapid reconstruction phylogenetic tree. *FSKD*.
- Meng, Z., Xiao, X., Li, J., Zhou, Y., Cao, W., & Shen, G. (2012) Cloud-GSQCT: a parallel approach to screen gene sequences for phylogenetics analysis. *2012 International Conference on Computer Science and Information Processing*, pp 660–663.

- Meng, Z., Zhou, Y., Li, J., Gao, Y., & Shen, Z. (2012) iDNABar: a rapid species identification toolbox for DNA Barcoding collection, preservation, identification and tracing. *Advances in Intelligent and Soft Computing* 124, pp 337–343.
- Pearse, W. D., Purvis, A., & Paradis, E. (2013) phyloGenerator: an automated phylogeny generation tool for ecologists. *Methods in Ecology and Evolution* 4(7), pp 692–698.
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2010) FastTree 2: Approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3), p e9490.
- Ratnasingham, S., & Hebert, P. D. N. (2007) BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Molecular Ecology Notes* 7(3), pp 355–364.
- Rindone, W. P. (1983) Genbank. *Trends in Pharmacological Sciences* 4(8), pp 326–326.
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012) MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology* 61(3), pp 539–542.
- Sanderson, M. J., Boss, D., Chen, D., Cranston, K. A., & Wehe, A. (2008) The PhyLoTA browser: Processing GenBank for molecular phylogenetics research. *Systematic Biology* 57(3), pp 335–346.
- Schindel, D., & Miller, S. E. (2005) DNA barcoding a useful tool for taxonomists. *Nature*.
- Smith, S. A., Beaulieu, J.M., & Donoghue, M.J. (2009) Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMCEvolutionary Biology* 37(9).
- Stamatakis, A. (2006) RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21), pp 2688–2690.
- Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatic*, 23(13), pp i549–i558.
- Webb, C. O., Ackerly, D. D., McPeck, M. A., & Donoghue, M. J. (2002) Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics* 33(1), pp 475–505.
- Wu, Z. Y., Lu, A. M., Tang, Y. C., Chen, Z. D., Li, D. Z. (2003) The Families and Genera of Angiosperms in China- A Comprehensive Analysis. *Science Press*.
- Zmasek, C. M., & Eddy, S. R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* 17, pp 383–384.

How to cite this article: Meng, Z, Dong, H, Li, J, Chen, Z, Zhou, Y, Wang, X and Zhang, S 2015 Darwintree: A Molecular Data Analysis and Application Environment for Phylogenetic Study. *Data Science Journal*, 14: 10, pp.1-10, DOI: <http://dx.doi.org/10.5334/dsj-2015-010>

Published: 22 May 2015

Copyright: © 2015 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License (CC-BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/3.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 