

Phylogenetics

jPHYDIT: a JAVA-based integrated environment for molecular phylogeny of ribosomal RNA sequences

Yoon-Seong Jeon¹, Hwanwon Chung¹, Sunyoung Park², Inae Hur¹,
Jae-Hak Lee¹ and Jongsik Chun^{1,2,*}

¹Interdisciplinary Program in Bioinformatics and ²School of Biological Sciences, Seoul National University, Shilim-dong, Kwanak-gu, Seoul, Republic of Korea

Received on December 9, 2004; revised on April 2, 2005; accepted on April 21, 2005

Advance Access publication April 26, 2005

ABSTRACT

Summary: jPHYDIT is a Java application designed to furnish a visual and integrated environment for molecular phylogeny. The program can be used to visualize intra-strand base-pairing information in secondary and tertiary structures of ribosomal RNA (rRNA) sequences. A function for the semi-automated alignment was included to facilitate handling of the database containing a large number of multiple-aligned rRNA sequences. Integration of nucleotide sequence editing, pairwise alignment, multiple alignment and phylogenetic treeing functions provide an easy and efficient way of analyzing rRNA sequences for molecular evolution, systematics, epidemiology and ecology.

Availability: jPHYDIT is available freely over the Internet at <http://chunlab.snu.ac.kr/jphydit/>. The platform-independent JAVA technology provides distributions for various operating systems and hardware architectures.

Contact: jchun@snu.ac.kr

INTRODUCTION

Owing to specific characteristics, such as its ubiquity, size and low evolutionary rates, ribosomal RNA (rRNA) has been a central framework for modern systematic, phylogenetic and ecological studies. Because of its popularity as a molecular marker, a large number of rRNA sequences were determined and are available from thousands of different species. For example, the number of GenBank nucleotide entries containing the term '16S rRNA' alone is over 178 000 as of 2004.

The correct alignment of multiple nucleotide sequences is a prerequisite for accurate and reliable phylogeny. rRNA secondary and tertiary structural models can be used to significantly improve the quality of multiple alignments, and their use has been recommended for phylogenetic analysis using rRNA sequences (Hickson *et al.*, 1996; Ludwig and Schleifer, 1994). Analysis of rRNA sequences involves multiple alignment of a large number of sequences; the task cannot be easily achieved using the currently available multiple alignment algorithms due to extensive computing costs. The multiple alignments provided by several rRNA sequence database sites such as the ribosomal database project (RDP) (Cole *et al.*, 2005) and European ribosomal RNA database (Wuyts *et al.*, 2004) are frequently used. However, these alignments are large (>10 000

sequences) and complex, and the addition of new sequences is a demanding task—for the database curators in laboratories or for individuals who wish to align new sequences using existing aligned sequences.

The approach commonly used in many laboratories, includes (1) to obtain the aligned sequences from a database; (2) to add new sequences to the existing alignment manual by or using semi-automated methods; and (3) to adjust manually using rRNA secondary structure information. jPHYDIT is a visual nucleotide sequence editor dedicated to such a process. Our program has an advantage over other RNA alignment programs, including DCSE (De Rijk and De Wachter, 1993) and MARNA (<http://www.bio.inf.uni-jena.de/Software/MARNA/>), as it is run on multiple operating systems and visualizes RNA secondary structural information during the editing process.

IMPLEMENTATION

jPHYDIT was written using JAVA but can be run on any operating system installed with a JAVA runtime environment 1.4 or higher. Since the aim is to provide an intuitive graphical user interface (GUI), the X-window system is required if jPHYDIT is run under Linux. Our program used multiple document interface and modular structure for the implementation of additional algorithms for the future. All computational approaches or algorithms required to carry out every subtask of molecular phylogeny were equipped according to this modularity strategy.

Semi-automated pairwise alignment and alignment editor displaying RNA pairing information

jPHYDIT is an integrated graphical environment which contains sequence editor function for manual adjustment using a simple and intuitive user interface. The optimal linear space pairwise alignment algorithm (Myers and Miller, 1988) was adapted for the semi-automated alignment process and can be used to append and align a new sequence to existing multiple alignments. A nucleotide sequence from any database holding pre-aligned sequence data, such as RDP, can be used as a template sequence. The resultant machine-driven alignment should be adjusted manually. By using this semi-automated alignment, the number of required manual adjustments is substantially reduced compared with template-free manual pairwise alignment.

*To whom correspondence should be addressed.

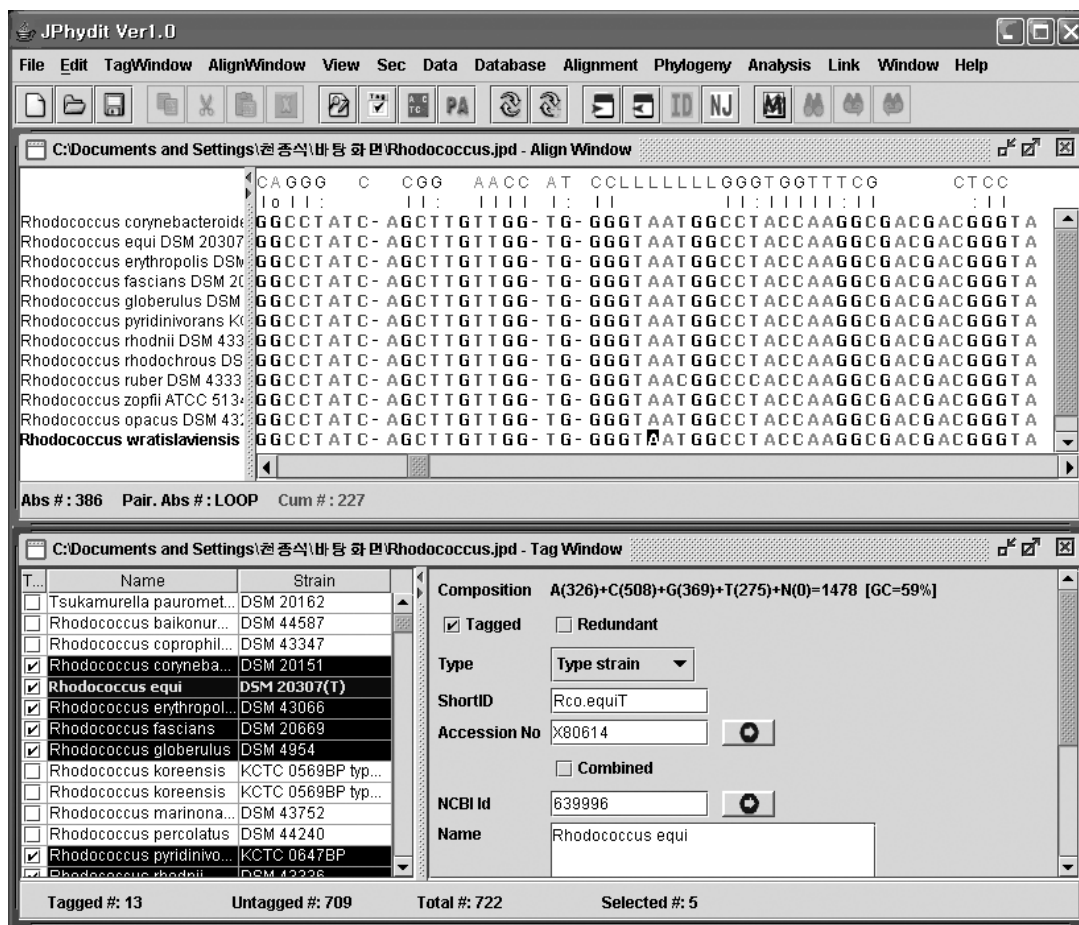


Fig. 1. The GUI of jPHYDIT. A sequence (bottom sequence in the upper half) can be imported from the GenBank and added to the existing multiple alignment using a semi-automated method. The nucleotide sequence editor displays intra-strand base pairing (top of alignment), which can be used in manual adjustment. L indicates the loop structure. The window for editing general information of a sequence is shown in the lower half.

The information for base pairings in RNA secondary and tertiary structures can be obtained from a multiple alignment and stored in a text file, called secondary structure file. jPHYDIT has a module for graphically displaying the stem structure along with sequence alignment, which assists users in manual adjustments (Fig. 1). Intra-strand base pairing information was obtained from the Comparative RNA website (Cannone *et al.*, 2002).

After carrying out the manual adjustment operations, the quality of sequencing and alignment can be evaluated on the basis of the secondary structure. The sequence quality value, Q , is calculated using the following equation:

$$Q = (P/T) \times 100,$$

where P denotes the number of matched position and T denotes the total number of intra-strand paired position.

Integrated environment for phylogeny

The goal of jPHYDIT is to provide all functions for molecular phylogenetic analysis. The current version contains nucleotide sequence editor, pairwise alignment, multiple alignment and neighbor-joining

treeing method (Saitou and Nei, 1987). Viewing and manipulation of the resultant phylogenetic trees are done by using other applications, such as TreeView (Page, 1996). The future version will include additional phylogeny inference methods, such as maximum parsimony.

Database management system for large number of sequences

Users can manage and share a large number of sequences using jPHYDIT's database module. It utilizes JDBC and MySQL for Internet-based multiuser environment. The system is suitable and useful for inter-laboratory projects, in which users in different locations need to share sequence data in real time.

ACKNOWLEDGEMENTS

This study was supported by a grant (01-PJ11-PG9-01BT00B-03) from the International Mobile Telecommunications 2000 R&D Project, Ministry of Information and Communication, Republic of Korea. S.P. was supported by a BK21 Fellowship from the Ministry of Education and Human Resources Development.

REFERENCES

- Cannone,J.J. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
- Cole,J.R. *et al.* (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.*, **33**(Database issue), D294–D296.
- De Rijk,P. and De Wachter,R. (1993) DCSE, an interactive tool for sequence alignment and secondary structure research. *Comput. Appl. Biosci.*, **9**, 735–740.
- Hickson,R.E. *et al.* (1996) Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. *Mol. Biol. Evol.*, **13**, 150–169.
- Ludwig,W. and Schleifer,K.H. (1994) Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol. Rev.*, **15**, 155–173.
- Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *Comput. Appl. Biosci.*, **4**, 11–17.
- Page,R.D. (1996) TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.*, **12**, 357–358.
- Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
- Wuyts,J. *et al.* (2004) The European ribosomal RNA database. *Nucleic Acids Res.*, **32**(Database issue), D101–D103.