Submitted to *Bioinformatics* for the publication as APPLICATION NOTES

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

Yoon-Seong Jeon 1, Hwanwon Chung 1, Sunyoung Park 2, Inae Hur 1, Jae-Hak Lee 1 and Jongsik Chun 1,2,*

1 Interdisciplinary Program in Bioinformatics and 2 School of Biological Sciences, Seoul National University, Shilim-dong, Kwanak-gu, Seoul, Republic of Korea, Tel: +82-2-880-8153, Fax: +82-2-888-4911, E-mail: jchun@snu.ac.kr

Running title: JAVA environment for rRNA gene phylogeny

*To whom correspondence should be addressed
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 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

Due to specific characteristics such as its ubiquity, size and low evolutionary rates, the 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 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 the prerequisite for the 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 the phylogenetic analysis using rRNA sequences (Hickson et al., 1996; Ludwig & Schleifer, 1994). Analysis of rRNA sequences involves multiple alignment of a large number of sequences; the task cannot be easily done using the currently available multiple alignment algorithms due to the extensive computing cost. 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 (Wutys et al., 2004) are frequently used. However, these alignments are large (> 10,000 sequences) and complex, and the addition of new sequences is demanding task either for the database curators in each laboratory or for individuals who wish to align new sequences with existing aligned sequences.

The common approach used in many laboratories includes: (i) to obtain the aligned sequences from a database; (ii) to add new sequences to the existing alignment by manual or semi-automated manner; and (iii) 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 & De Wachter, 1993) and MARNA ([http://www.bio.inf.uni-jena.de/Software/MARNAs](http://www.bio.inf.uni-jena.de/Software/MARNAs)) as it is run on multiple operating systems and visualizes RNA secondary structural information while editing process.

**IMPLEMENTATION**

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

**Semi-automated pairwise alignment and Alignment editor with displaying RNA pairing information**

jPHYDIT is an integrated graphical environment containing sequence editor function for manual adjustment with simple and intuitive user interface. The optimal linear space pairwise alignment algorithm (Myers & Miller, 1988) was adapted for the semi-automated alignment process which can be used to append and align a new sequence to existing multiple alignment. 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 to template free manual pairwise alignment.

The information for base pairings in RNA secondary and tertiary structure 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 (Figure 1). Intra-strand base pairing information was obtained from the Comparative RNA Web Site (Cannone et al., 2002).

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

\[ Q = \left( \frac{P}{T} \right) \times 100, \]

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 & 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 the different locations need to share the sequence data in real time.

ACKNOWLEDGEMENTS

This study was supported by a grant (01-PJ11-PG9-01BT00B-0003) of the
International Mobile Telecommunications 2000 R&D Project, Ministry of Information &
Communication, Republic of Korea. SP was supported by a BK21 Fellowship from the
Ministry of Education and Human Resources Development.

REFERENCES

web (CRW) site: an online database of comparative sequence and structure
information for ribosomal, intron, and other RNAs. *BMC Bioinformatics, 3*, 2.
tools for high-throughput rRNA analysis. *Nucleic Acids Res.*, 33 Database Issue,


Figure 1. The graphical user interface of jPHYDIT. A sequence (bottom sequence in the upper half) can be imported from Genbank and added to the existing multiple alignment using a semi-automated manner. The nucleotide sequence editor displays intra-strand base paring (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.