EzEditor: a versatile sequence alignment editor for both rRNA- and protein-coding genes

Yoon-Seong Jeon,1,2 Kihyun Lee,3 Sang-Cheol Park,1 Bong-Soo Kim,2 Yong-Joon Cho,2 Sung-Min Ha3 and Jongsik Chun1,2,3

1Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
2ChunLab, Inc., Seoul National University, Seoul, Republic of Korea
3School of Biological Sciences, Seoul National University, Seoul, Republic of Korea

EzEditor is a Java-based molecular sequence editor allowing manipulation of both DNA and protein sequence alignments for phylogenetic analysis. It has multiple features optimized to connect initial computer-generated multiple alignment and subsequent phylogenetic analysis by providing manual editing with reference to biological information specific to the genes under consideration. It provides various functionalities for editing rRNA alignments using secondary structure information. In addition, it supports simultaneous editing of both DNA sequences and their translated protein sequences for protein-coding genes. EzEditor is, to our knowledge, the first sequence editing software designed for both rRNA- and protein-coding genes with the visualization of biologically relevant information and should be useful in molecular phylogenetic studies. EzEditor is based on Java, can be run on all major computer operating systems and is freely available from http://sw.ezbiocloud.net/ezeditor/.

The general procedure for molecular phylogenetic analysis consists of two major steps, namely multiple sequence alignment and reconstruction of phylogenetic trees. The former is regarded as important as it can affect the accuracy of all downstream analyses, including the latter. Many computer programs are available for multiple alignment of DNA or protein sequences, with the CLUSTAL-series programs being the most popular (Thompson et al., 2002). However, computer-generated sequence alignment is often required to be improved by considering biological knowledge such as secondary structure of RNA and the reading frames of protein-coding genes. This task can be carried out using molecular sequence editing software, called sequence editors, by which computer-generated alignments are viewed visually and adjusted manually by adding or deleting gaps.

Several sequence editors are available for general usage, including SEAVIEW (Galtier et al., 1996), BioEdit (Hall, 1999), DNAAlignEditor (Sanchez-Villeda et al., 2008), INTERALIGN (Pible et al., 2005) and JalView (Waterhouse et al., 2009). In the case of rRNA genes, which are the most widely used phylogenetic markers, special sequence editing programs have been developed, including ARB (Ludwig et al., 2004) and jPHYDIT (Jeon et al., 2005). These software tools allow users to consider secondary structure information during manual editing.

Conserved protein-coding genes, such as rpoB, recA and gyrB, are also widely used in molecular phylogenetic studies (Case et al., 2007; Feng et al., 1997). Unlike rRNA genes, protein-coding genes can be used either as DNA or translated protein sequences in phylogenetic analyses. Because of the degenerate nature of the genetic code, single amino acids can be encoded by multiple codons, and DNA sequences in coding regions contain position-specific information as a component of codons, depending on the reading frame. Therefore, in the case of amino acid sequences, the original DNA sequences coding for proteins are better aligned by codon-based alignment (Goldman & Yang, 1994) in which protein sequences are first aligned and their DNA sequences are then rearranged by inserting gaps, on the basis of the previously aligned protein sequences. In this scheme, one gap in the protein sequence alignment is translated into three consecutive gaps in the DNA sequence alignment. There are a few software tools and web services to achieve this task, including PAL2NAL (Suyama et al., 2006), RevTrans (Wernersson & Pedersen, 2003) and transAlign (Bininda-Emonds, 2005). However, to our knowledge, there is no sequence editor that allows simultaneous editing of both DNA sequences and their translated protein sequences for protein-coding genes. Here, we introduce a new sequence editor, named EzEditor, for simultaneous codon-based editing of protein and DNA sequence alignments. Since it is a descendant of jPHYDIT (Jeon et al., 2005) it provides all the functionality for editing rRNA alignment using secondary structure information.

EzEditor is written in the JAVA language and can be run on the Java Runtime Environment of all major operating systems. The key feature of EzEditor is the dual editing
Fig. 1. Two different editing modes of EzEditor. Screen-shots of (a) protein-coding sequence alignment in which both DNA and translated protein sequences are simultaneously displayed for manual editing and (b) DNA sequence alignment for 16S rRNA genes showing RNA secondary structure at the top of the window.
function that allows simultaneous display and editing of DNA and corresponding translated protein sequence alignments (Fig. 1a). Adding or deleting a single gap in the protein sequence alignment window will result in simultaneous addition or deletion of three gaps in the DNA sequence alignment window, respectively. EzEditor uses external software, such as CLUSTAL W (Thompson et al., 2002), for computer-assisted multiple alignment. The edited alignment can be exported to various types of phylogenetic software, including MEGA (Tamura et al., 2007), PHYLIP (Felsenstein, 1993) and PAUP (Swofford, 2002). Using this codon-based DNA alignment, users can obtain multiple sequence alignments containing only the first and second positions of each codon. Various types of sequence data can be imported into EzEditor, including data in the FASTA and GenBank formats. Visualization of RNA secondary structure within the editing screen and various key manipulations for editing rRNA gene alignments are improved in comparison with the previous version, i.e. jPHYDIT (Jeon et al., 2005) (Fig. 1b). EzEditor represents the first versatile sequence editor, to our knowledge, to handle both rRNA- and protein-coding genes, which is useful for molecular phylogenetic analysis.

Acknowledgements

This work was supported by the Korean Ministry of Knowledge Economy (MKE)-funded Industrial Strategic Technology Development Program (10040176).

References


