Artemis: sequence visualization and annotation

Kim Rutherford¹, Julian Parkhill¹, James Crook², Terry Horsnell², Peter Rice¹, Marie-Adèle Rajandream¹ and Bart Barrell¹

¹The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK and ²The MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK

Received on March 22, 2000; accepted on May 30, 2000

Abstract

Summary: Artemis is a DNA sequence visualization and annotation tool that allows the results of any analysis or sets of analyses to be viewed in the context of the sequence and its six-frame translation. Artemis is especially useful in analysing the compact genomes of bacteria, archaea and lower eukaryotes, and will cope with sequences of any size from small genes to whole genomes. It is implemented in Java, and can be run on any suitable platform. Sequences and annotation can be read and written directly in EMBL, GenBank and GFF format.

Availability: Artemis is available under the GNU General Public License from http://www.sanger.ac.uk/Software/Artemis

Contact: kmr@sanger.ac.uk

Visualization of features within a sequence is an essential component of modern molecular biology. Some commercial programs exist to meet this need, but most freely available programs are either designed for large eukaryotic genomes (such as AceDB; http://www.acedb.org/) or have limited capabilities. We have developed Artemis, a Java-based sequence visualization tool, for our own in-house analysis. It is used extensively in our laboratories for the annotation of bacterial and lower eukaryotic genomes, most recently for the annotation of the complete genomes of Campylobacter jejuni and Neisseria meningitidis, and the chromosomes of Plasmodium falciparum. Artemis grew out of the concepts developed in DIANA (for DIsplay and ANAlyse; J.C., T.H. & P.R., unpublished), which was previously used at the MRC-LMB and the Sanger Centre for sequence analysis.

Artemis can be used simply as a sequence viewer, on any platform running Java (Unix-, Macintosh- or PC-based), and will allow the visualization of sequence and annotation taken directly from EMBL (Baker et al., 2000) and GenBank (Benson et al., 2000) format files. Properties of the sequence, such as G+C content, G/C skew (Lobry, 1996), frame-specific G+C content (Bibb et al., 1984), codon usage etc. can be directly plotted against the sequence and its features. Each plot allows dynamic modification of the window size used for the calculation, and the sequence and plots can be zoomed together into the single base level or out to the complete genome. Two sequence windows exist which can be used to view the same sequence at different zoom levels simultaneously. Properties of individual protein features, such as hydrophobicity and hydrophilicity can also be viewed in the same way. Artemis can also be run as a WWW applet, allowing the serving of sequence and annotation over the web in an interactive format.

Artemis is specifically designed to be used as an annotation tool. In this mode, the results of any external analysis program can be parsed into the correct format (EMBL/GenBank feature table format; http://www.ebi.ac.uk/embl/Documentation/User_manual/format.html) and be overlaid on the sequence. These features can then be modified and moved between files to build up a single annotation table. External programs, such as FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990) can be run on any number of individual features, and the results viewed from within the program. The qualifiers attached to each feature can be edited directly, and users can add any number of custom qualifiers. Because Artemis is designed to be as portable as possible, and to display the results of as wide a variety of analyses as possible, complex algorithms have not been built into the program itself. Sample parsing scripts for commonly used programs are available as part of the distribution.

Acknowledgements

The development of Artemis was funded by the Wellcome Trust’s Beowulf Genomics initiative through its support of the Sanger Centre Pathogen Sequencing Unit. The early development of DIANA was funded by the MRC through its support of the LMB.

References


944 © Oxford University Press 2000
Fig. 1. Top left; the main Artemis window showing a section of the *Streptomyces coelicolor* cosmid EMBL:SCD82. The upper window is a frame-specific G+C plot, the second is the main sequence display with three forward reading frames, forward and reverse DNA lines separated by a scale bar, and three reverse reading frames. Vertical bars represent stop codons, open boxes show features. The third window is the secondary sequence display zoomed into the nucleotide level. The bottom window is a feature list. Top right; the sequence overview, showing the results of global calculations on the sequence. Middle right; protein property plots for one coding sequence. Bottom right; the feature edit window: Artemis makes all legitimate keys and qualifiers available from drop-down menus. Bottom; G+C plot and main sequence display for a highly spliced gene from *Plasmodium falciparum* chromosome 4 (EMBL:PFMAL4P2).


