BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 16 2005, pages 3422–3423
doi:10.1093/bioinformatics/bti553

Genome analysis

ACT: the Artemis comparison tool

Tim J. Carver[1,*], Kim M. Rutherford[2], Matthew Berriman[1], Marie-Adele Rajandream[1], Barclay G. Barrell[1] and Julian Parkhill[1]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK and [2]Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK

Received on April 20, 2005; revised on June 7, 2005; accepted on June 22, 2005
Advance Access publication June 23, 2005

ABSTRACT

The Artemis Comparison Tool (ACT) allows an interactive visualisation of comparisons between complete genome sequences and associated annotations. The comparison data can be generated with several different programs; BLASTN, TBLASTX or Mummer comparisons between genomic DNA sequences, or orthologue tables generated by reciprocal FASTA comparison between protein sets. It is possible to identify regions of similarity, insertions and rearrangements at any level from the whole genome to base-pair differences. ACT uses Artemis components to display the sequences and so inherits powerful searching and analysis tools. ACT is part of the Artemis distribution and is similarly open source, written in Java and can run on any Java enabled platform, including UNIX, Macintosh and Windows.

Availability: ACT is freely available (under a GPL licence) for download from the Sanger Institute web site, http://www.sanger.ac.uk

Contact: artemis@sanger.ac.uk

Comparative genomics is an increasingly important step in the annotation and analysis process, allowing phenotypic differences between strains and species to be correlated with changes in the chromosomes. Since 2001 (Cole *et al.*, 2001) ACT has been used and developed as a tool to carry out pair-wise genome comparison. Eukaryotic and prokaryotic chromosomes of five or more megabases can be comfortably viewed on a desktop system, with larger comparisons possible on larger systems. ACT is portable and can be run on UNIX, Windows and MacOSX.

Sequences can be read in EMBL, Genbank, GFF and FASTA formats. Comparison data can be generated directly by BLAST, or parsed from other comparison tools. For a pair of sequences, one is designated the query sequence and the other the subject sequence i.e. the database. The sequences are aligned with the subject sequences above the query sequence, showing the features at the nucleotide and amino acid level. The sequences are joined by coloured bands that represent the matching regions, enabling the user to intuitively visualise and gain biological insight from the comparison.

ACT can directly read the output of BLAST version 2.2.2 or higher generated by BLASTn or tBLASTx or MEGABLAST. Web sites exist (Double ACT and WebACT, www.webact.org) where sequences can be pasted in to generate the comparison files. Alternatively, the output of other comparisons (WUBLAST, BLAST1.4, Mummer, reciprocal best match) can be parsed into a simple one-line-per-match format, and used by ACT.

*To whom correspondence should be addressed.

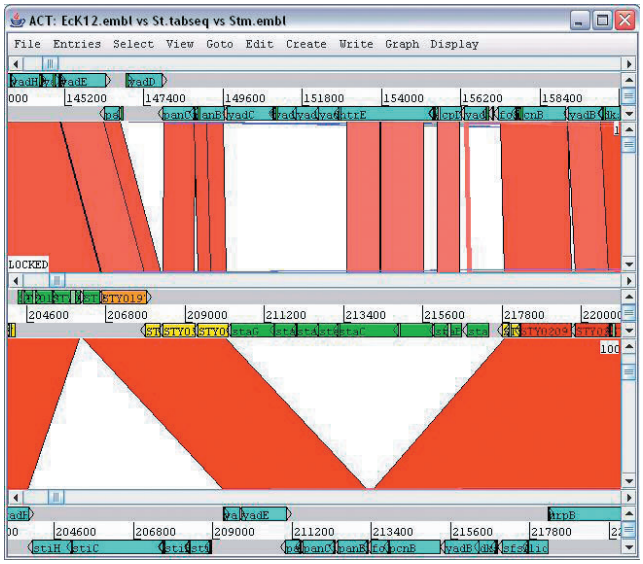

Fig. 1. BLASTN comparison of part of three sequences: *Escherichia coli* K12, *Salmonella* Typhi CT18 and *Salmonella* Typhimurium LT2 (from top to bottom). This graphically shows the relationship between each pair of sequences. The displays are collapsed to show all forward and reverse features on single lines (coloured boxes represent coding sequences). Insertions of *sta* and *sti* genes in the two *Salmonellae* can be clearly seen, as can the reduced similarity (lighter and absent match lines) between the *yad*/*htrE* genes in *E.coli* and the *sta* genes in *S.*Typhi.

The red and blue bands represent the forward and reverse matches, respectively. ACT can make reverse matches easier to view by flipping either of the sequences round so that they are seen in register. The intensity of the colour bands is proportional to the percent identity of the match, within the range chosen, giving a visual indication of the strength of the matches. Double clicking on one of these bands will centre the associated matching regions in each sequence, and the sequences can be scrolled individually or together.

A highlighted match region turns yellow and information about that BLAST hit is shown, e.g. the score and percentage identity. To view, or scroll through, all the matches that overlap a selected feature or region, the 'view selected matches' function can be used. The level of similarities displayed can be adjusted to increase or decrease sensitivity. Matches can be filtered based on their length, percentage identity or score, and only the filtered hits are then displayed by ACT.

ACT re-uses the Artemis (Rutherford *et al*., 2000 and Berriman *et al*., 2003) code to display the sequence and features. As a result it inherits a number of the tools and features of Artemis. Graphs based on numerous DNA properties (GC content, GC bias, codon usage, dinucleotide frequency etc.) can be displayed for each sequence, and will zoom and scroll with them. Hydophobicity (Kyte and Doolittle, 1982), hydrophilicity (Hopp and Woods, 1981) and coiled coils (Lupas *et al*., 1991) plots can be shown for protein features. The 'feature selector' and 'navigator' are powerful tools in moving around or searching for features based on various criteria, e.g. location, feature name, length, position and amino acid motif contained. Sequences can be opened in Artemis during an ACT session and edited, and the changes appear simultaneously in ACT.

In ACT, several sequences can be compared and analysed by stacking the multiple pair-wise comparisons (Fig. 1). Each sequence becomes a query and subject, except those at the boundaries. The number of sequences is limited only by the size of the screen and the available memory but three, four and even up to seven pair-wise comparisons have been used.

Other features of ACT include the ability to study regions of difference. It is possible to create features from non-matching regions or select features that lie within these regions. ACT images (screenshots) can be created and saved to PNG or JPEG files for publication.

Conflict of Interest: none declared.

REFERENCES

Berriman,M. and Rutherford,K. (2003) Viewing and annotating sequence data with Artemis. *Brief. Bioinformatics*, **4**, 124–132.

Cole,S.T. *et al*. (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*., **157**, 105–132.

Lupas,A. *et al*. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.

Rutherford,K. *et al*. (2000) Artemis: sequence visualisation and annotation. *Bioinformatics*, **16**, 944–945.