BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 18 2005, pages 3665–3666
doi:10.1093/bioinformatics/bti601

Genome analysis

WebACT—an online companion for the Artemis Comparison Tool

J. C. Abbott[1,*], D. M. Aanensen[2], K. Rutherford[3,†], S. Butcher[1] and B. G. Spratt[2]

[1]Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, UK, [2]Department of Infectious Disease Epidemiology, Imperial College London, London, W2 1PG, UK and [3]Pathogen Sequencing Unit, Sanger Institute, Cambridge CB10 1SA, UK

Received on May 23, 2005; revised and accepted on July 27, 2005
Advance Access publication August 2, 2005

ABSTRACT

Summary: WebACT is an online resource which enables the rapid provision of simultaneous BLAST comparisons between up to five genomic sequences in a format amenable for visualization with the well-known Artemis Comparison Tool (ACT). Comparisons can be generated on-the-fly using sequences directly retrieved via EMBL database queries, or by entering or uploading user sequences. Furthermore, pre-computed comparisons are available between all publicly available, completed prokaryotic genomes and plasmids currently contained within the Genome Reviews database (372 sequences, representing 175 different species). The system is designed to minimize the volume of downloaded data and maximize performance. Genome sequences, annotation and pre-computed comparisons are stored in a relational database allowing flexible querying based on user-defined sequence regions, from whole genome to a defined region flanking a specified gene. Comparison and sequence files, whether computed online or retrieved from the database of pre-computed genome comparisons, can be viewed online using ACT and are available for download.

Availability: Freely accessible at http://www.webact.org

Contact: admin@webact.org

Supplementary information: User guide and worked examples are available at http://www.webact.org/WebACT/docs

INTRODUCTION

The Artemis Comparison Tool (ACT) is a graphical DNA sequence comparison viewer allowing the results of a BLASTN or TBLASTX search to be viewed between sequences of interest, while highlighting available annotation (Carver *et al*., 2005). Presently, the generation of suitable sequence comparison files and their subsequent loading into ACT is the responsibility of the user. ACT requires the input of pre-generated comparison files in either the tab-delimited output format of BLAST (Altschul *et al*., 1997) or MSPCrunch (Sonnhammer and Durbin, 1994), together with the original sequences and their annotations, in EMBL or GenBank formats. For the uninitiated bench scientist, access to the necessary data, computational resources and the experience to generate these files efficiently is currently a significant obstacle to the usage of ACT.

WebACT is an online resource providing BLAST comparison and sequence files in appropriate formats for ACT, allowing the generation of comparisons based on sequences contained within the EMBL database (Kanz *et al*., 2005), from user submitted sequences, or selected from a database of pre-computed comparisons. Provision of pre-computed comparisons in this manner results in a significantly faster turn-round of prokaryotic sequence queries. Worked examples demonstrating use of WebACT are available alongside the documentation on the WebACT site.

PRE-BUILT SEQUENCE COMPARISONS

Sequences for pre-computed comparisons are sourced from the Genome Reviews database (Kersey *et al*., 2005). This database was used in preference to the original genome entries in the EMBL/GenBank database as it contains only completed sequences, which have more consistent annotation in a format appropriate to the requirements of ACT.

Pre-computed comparisons were generated using NCBI BLAST, with the results obtained in tab-delimited format. BLAST comparisons were carried out using a word size of nine and soft DUST masking. Each sequence was initially formatted as a BLAST database, which was also chunked into 100 kb segments with a 1 kb overlap for use as a query sequence. This approach avoids some of the problems inherent in running BLAST against long gene-rich sequences (e.g. Schwartz *et al*., 2000; Korf *et al*., 2003). An all-against-all set of pairwise BLASTN comparisons (including self-comparisons) was generated in such a manner that each comparison was only carried out once. All analysis results were parsed and stored in the WebACT database.

Up to five sequences can be selected from the database for comparative display. WebACT allows selection of the full-length sequence (i.e. an entire bacterial chromosome, or plasmid), a defined set of base co-ordinates, or a named gene and specified length of flanking sequence in the corresponding genome(s). Gene names can be either manually entered or chosen from a list of known genes present within the annotated genomes. Selection of the genomic region to be displayed can be made by applying the same criteria to each of the sequences, or defining each region individually. The comparison data

*To whom correspondence should be addressed.

†Present address: Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK

can be screened on the basis of the BLAST *E*-value before loading into ACT.

The sequences stored in the WebACT database are compared with those made available by the EBI on a monthly basis, and the sequence, annotation and comparisons are automatically updated as appropriate. Newly released sequences are also incorporated into the database at this time. Sequence records are parsed using Bioperl (Stajich *et al*., 2002), stripped of features and sequence data and stored in the database as serialized objects. The sequence is stored in 100 kb chunks, while each sequence feature is stored as a serialized object along with its genome co-ordinates, permitting the rapid assembly of a sequence record representing the requested region. Full sequence records are also stored as compressed flat-files to optimize performance when such records are requested.

EMBL QUERIES AND USER-DEFINED SEQUENCES

Comparisons can also be generated on-the-fly from user-defined sequences. These may be uploaded in EMBL, GenBank or FASTA formats, or selected by EMBL accession number. For accession numbers relating to the Contig division of the EMBL database, each constituent record is retrieved and automatically assembled into a record containing the full set of CDS features for the selected sequence. BLAST comparisons between the chosen sequences are then carried out in a pairwise manner. Generation of large sequence comparisons is computationally expensive, consequently completion of the comparisons can be notified via email.

VISUALIZATION AND DATA DOWNLOAD

Once a comparison has been generated, ACT can be launched directly from the user's web browser using Java Web Start. All results are retained on the server for 7 days; however, a WebACT session can be downloaded as a single file and retained by the user for use off-line. The downloaded file can be reloaded into WebACT at any point, from which the comparison may be viewed without time-consuming regeneration. An additional advantage of using the pre-computed comparisons is that a small file (∼2 kb), which defines the user's sequence selections, can be downloaded and it allows WebACT to reconstruct the comparison at a later date, removing the need for the user to download large quantities of data. Sequence data downloaded from the WebACT database can also be reloaded to permit a comparison to be repeated using a different algorithm or set of parameters.

IMPLEMENTATION

WebACT is a mod_perl application built using the CGI:::Application framework. Sequences, annotation and BLAST hits are stored in a MySQL relational database. Extensive use is made of the Bioperl modules (Stajich *et al*., 2002) for handling sequence data. User-submitted queries are managed through the Sun Grid Engine 6.0 job scheduler. Pre-computed comparisons were generated using an AMD Opteron based cluster running Red Hat Enterprise Linux 3.0, via Sun Grid Engine. WebACT has been tested using Internet Explorer, Firefox, Opera and Safari browsers running on Windows, Linux and Mac OSX. The only client-side requirements are for a supported JavaScript enabled web browser, and a Java 1.3 or newer installation. In order to launch the ACT application directly from WebACT, Java Web Start is required.

OUTLOOK

Plans for future development of WebACT include the addition of a facility to permit a comparison selected from the database to be re-run with user-defined parameters, and the incorporation of BLASTZ (Schwartz *et al*., 2003) as an additional comparison algorithm.

ACKNOWLEDGEMENTS

The authors are grateful to the London E-Science Centre (http://www.lesc.imperial.ac.uk) for access to high performance computing resources. This work was supported by the Faculties of Life Sciences and Medicine, Imperial College London and the Wellcome Trust.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Carver,T.J. *et al*. (2005) ACT: The Artemis Comparison Tool. *Bioinformatics*, in press.

Kanz,C. *et al*. (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **33**, D29–D33.

Kersey,P. *et al*. (2005) Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, **33**, 297–302.

Korf,I. *et al*. (2003) *BLAST*. O'Reilly and Associates, Sebastapol.

Sonnhammer,E.L.L. and Durbin,R. (1994) A workbench for large scale sequence homology analysis. *Comput. Appl. Biosci.*, **10**, 301–307.

Stajich,J.E. *et al*. (2002) The Bioperl Toolkit: Perl modules for the life sciences. *Genome Res.*, **10**, 1611–1618.

Schwartz,S. *et al*. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.*, **10**, 577–586.

Schwartz,S. *et al*. (2003) Human–mouse alignments with BLASTZ. *Genome Res*., **13**, 103–107.