Artemis: sequence visualization and annotation

Kim Rutherford[1], Julian Parkhill[1], James Crook[2], Terry Horsnell[2], Peter Rice[1], Marie-Adèle Rajandream[1] and Bart Barrell[1]

[1]The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK and [2]The 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 Artemis grew out of the concepts developed in DIANA (for DIsplay and ANAlyse; J.C., T.H. & P.R., *unpublished*), which was previously used at the MRC-LMB and the Sanger Centre for sequence analysis.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 a 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.

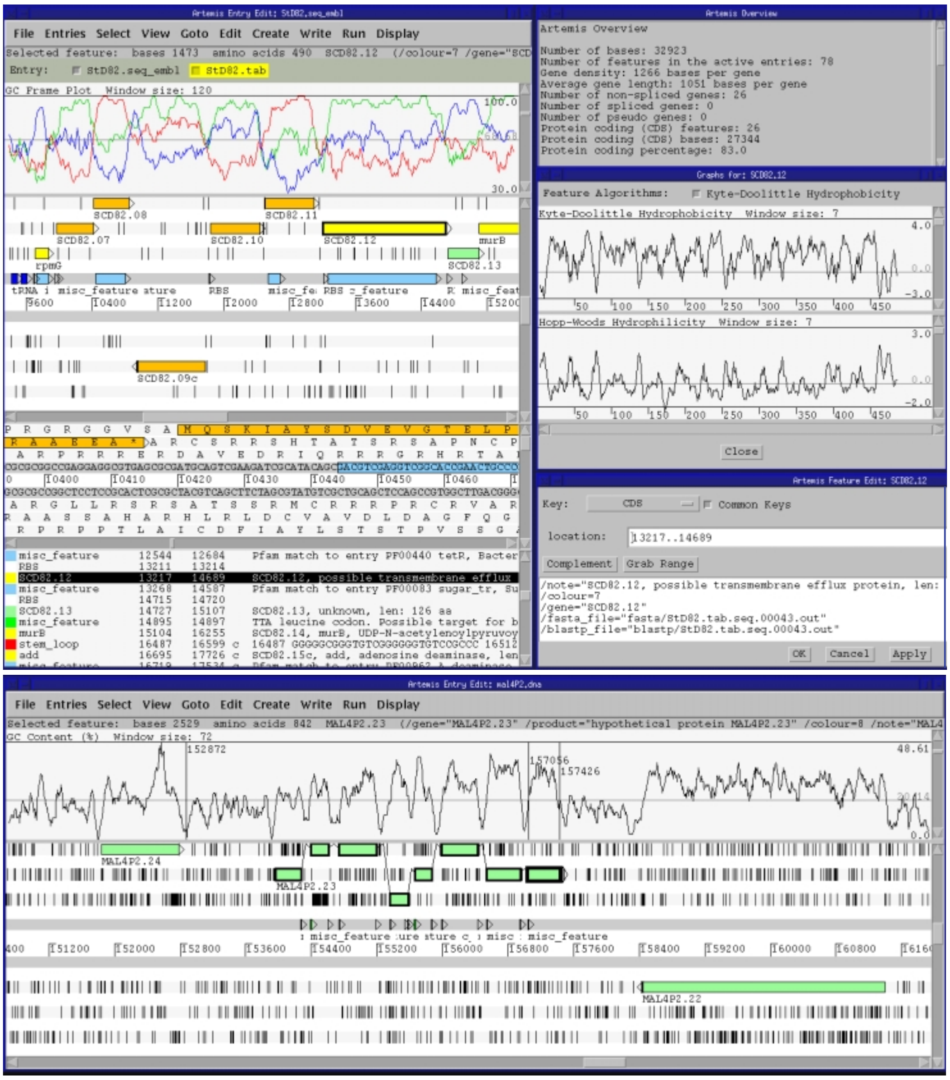
Acknowledgements

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

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

© Oxford University Press 2000



```
Selected feature:  bases 1473  amino acids 490  SCD82.12  (/colour=7 /gene="SCD
Entry:  ▨ StD82.seq_embl  ▨ StD82.tab
```

```
P  R  G  R  G  G  V  S  A  M  Q  S  K  I  A  Y  S  D  V  E  V  G  T  E  L  P
R  A  A  E  E  A  *  A  R  C  S  R  R  S  H  T  A  T  S  R  S  A  P  N  C  P
 A  R  P  R  R  R  E  R  D  A  V  E  D  R  I  Q  R  R  R  G  R  H  R  T  A
CGCGCGGCCGAGGAGGCGTGAGCGCGATGCAGTCGAAGATCGCATACAGCGACGTCGAGGTCGGCACCGAACTGCCC
0     10400     10410     10420     10430     10440     10450     10460     1
GCGCGCCGGCTCCTCCGCACTCGCGCTACGTCAGCTTCTAGCGTATGTCGCTGCAGCTCCAGCCGTGGCTTGACGGG
 A  R  G  L  L  R  S  R  S  A  T  S  S  R  M  C  R  R  R  P  R  C  R  V  A  R
R  A  A  S  S  A  H  A  R  H  L  R  L  D  C  V  A  V  D  L  D  A  G  F  Q  G
 R  P  R  P  P  T  L  A  I  C  D  F  I  A  Y  L  S  T  S  T  P  V  S  S  G
```

```
misc_feature   12544   12684     Pfam match to entry PF00440 tetR, Bacter
RBS            13211   13214
SCD82.12       13217   14689     SCD82.12, possible transmembrane efflux
misc_feature   13268   14587     Pfam match to entry PF00083 sugar_tr, Su
RBS            14715   14720
SCD82.13       14727   15107     SCD82.13, unknown, len: 126 aa
misc_feature   14895   14897     TTA leucine codon. Possible target for b
murB           15104   16255     SCD82.14, murB, UDP-N-acetylenoylpyruvoy
stem_loop      16487   16599 c   16487 GGGGGCGGGTGTCGGGGGGTGTCCGCCC 16512
add            16695   17726 c   SCD82.15c, add, adenosine deaminase, len
```

```
Artemis Overview

Number of bases: 32923
Number of features in the active entries: 78
Gene density: 1266 bases per gene
Average gene length: 1051 bases per gene
Number of non-spliced genes: 26
Number of spliced genes: 0
Number of pseudo genes: 0
Protein coding (CDS) features: 26
Protein coding (CDS) bases: 27344
Protein coding percentage: 83.0
```

```
Feature Algorithms:   ▨ Kyte-Doolittle Hydrophobicity
```

```
Key:   CDS   ▨ Common Keys
location:   13217..14689
/note="SCD82.12, possible transmembrane efflux protein, len:
/colour=7
/gene="SCD82.12"
/fasta_file="fasta/StD82.tab.seq.00043.out"
/blastp_file="blastp/StD82.tab.seq.00043.out"
```

```
Selected feature:  bases 2529  amino acids 842  MAL4P2.23  (/gene="MAL4P2.23" /product="hypothetical protein MAL4P2.23" /colour=8 /note="MAL4
```

Fig. 1. Top left; the main Artemis windoow 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).

Baker,W., van den Broek,A., Camon,E., Hingamp,P., Sterk,P., Stoesser,G. and Tuli,M.A. (2000) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **28**, 19–23.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.

Bibb,M.J., Findlay,P.R. and Johnson,M.W. (1984) The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene*, **30**, 157–166.

Lobry,J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665.

Parkhill,J., Wren,B.W., Mungall,K., Ketley,J.M., Churcher,C., Basham,D., Chillingworth,T., Davies,R.M., Feltwell,T., Holroyd,S., Jagels,K., Karlyshev,A.V., Moule,S., Pallen,M.J., Penn,C.W., Quail,M.A., Rajandream,M.A., Rutherford,K.M., van Vliet,A.H., Whitehead,S. and Barrell,B.G. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**, 665–668.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.