Viewing and annotating sequence data with Artemis

Matt Berriman and Kim Rutherford

Date received (in revised form): 27th February 2003

Matt Berriman
is a Senior Computer Biologist at The Wellcome Trust Sanger Institute, annotating the genomes of eukaryotic pathogens. He holds a PhD from the London School of Hygiene & Tropical Medicine and has previously worked at the University of Dundee and Rockefeller University.

Kim Rutherford
is a Principal Computer Programmer at The Wellcome Trust Sanger Institute, developing genome analysis and annotation software. He graduated from the Victoria University of Wellington, New Zealand.

Keywords: *Artemis, annotation, genome project, genome sequence, gene prediction, annotation software*

Abstract

Artemis is a widely used software tool for annotating and viewing sequence data. No database is required to use Artemis. Instead, individual sequence data files can be analysed with little or no formatting, making it particularly suited to the study of small genomes and chromosomes, and straightforward for a novice user to get started. Since its release in 1999, Artemis has been used to annotate a diverse collection of prokaryotic and eukaryotic genomes, ranging from *Streptomyces coelicolor* to, more recently, a large proportion of the *Plasmodium falciparum* genome. Artemis allows annotated genomes to be easily browsed and makes it simple to add useful biological information to raw sequence data. This paper gives an overview of some of the features of Artemis and includes how it facilitates manual gene prediction and can provide an overview of entire chromosomes or small compact genomes – useful for uncovering unusual features such as pathogenicity islands.

INTRODUCTION

Hundreds of small genomes have been completed or are in progress[1] and more are becoming available all the time. As technological improvements drive down sequencing costs the trend will continue. New projects are now feasible such as sequencing several related species or even strains or isolates from the same species. Artemis is a freely available program designed to help biologists view or annotate DNA sequence.[2] It is written in the Java programming language; hence it is portable across multiple computer platforms. In addition to the visualisation of sequence features, it allows the results of analyses within the context of the sequence and its six-frame translation to be examined. Although the software was originally designed as a tool for preparing and editing files in EMBL format for submission to the public databases, more recent versions contain several features that provide a framework for interrogating data from other flat file annotation formats. At The Wellcome Trust Sanger Institute, Artemis is now the main annotation tool for pathogen genomes and with more than 1,000 hits per week in 2002 to its download site, it must rank among the most highly used annotation tools.

GETTING STARTED WITH ARTEMIS

Separate versions of Artemis are available for Unix/Linux, Microsoft Windows or for Macintosh users. They are all free and can be obtained from the internet.[3] It is distributed under the GNU General Public Licence.[4] The internet site for downloading Artemis also contains detailed documentation for installing and running the software.

Artemis can read several file formats (including EMBL, GenBank, FASTA and GFF), thus files containing sequence data, with accompanying annotation or sequence data alone can be opened. In both cases annotation can be added to the file or further annotation can be 'layered' onto the sequence from separate files. Each layer of annotation is called an entry and in most cases consists simply of a table or 'Tab' file – a file consisting of only the feature table lines (those marked FT) from an EMBL entry. The use of entries is central to using Artemis effectively; it allows different types of annotation

M. Berriman,
The Wellcome Trust Sanger Institute,
Genome Campus,
Hinxton CB10 1SA, UK

Tel: +44 (0) 1223 494817
e-mail: mb4@sanger.ac.uk

representing different lines of evidence to be viewed independently.

Upon opening, the main Artemis edit window comprises a menu bar, an entries bar and panels showing three separate views of the sequence and annotation (Figure 1). The entries line contains one button for each entry that has been loaded. These buttons allow the user to set one entry as a default entry, where changes will be saved, and to activate or inactivate entries. Only those features from active entries are visible to the user. Much of the window is used for a summary overview of active features, such as annotated coding sequences, and a close-up view of the DNA sequence. Underneath is a textual summary, which by clicking on the right-hand mouse button (or the Apple key and the mouse button on Macintosh computers) can be set to show different aspects of the annotation, whether it be the feature type (ie EMBL feature keys such as `CDS`, `tRNA`, `misc_feature`, etc), gene name, product name or a description from another annotated field. Both the overview and DNA views show the DNA and its translation in all six reading frames and operate very similarly, allowing the user to zoom in and out or to scroll along a sequence. Freedom to change magnification levels allows details of individual genes to be examined. At maximum zoom very detailed annotation is possible, for instance in the fission yeast *Schizosaccharomyces pombe* individual donor and acceptor sequences for pre-mRNA splicing, which have been annotated across the entire genome,[5] can be viewed (Figure 1, middle panel). Conversely, a 'zoomed out' view allows many genes to be reviewed at once and can reveal patterns of gene clustering. This has helped resolve directional gene clusters, which may correspond to large

A textual summary can be set to show different aspects of the annotation

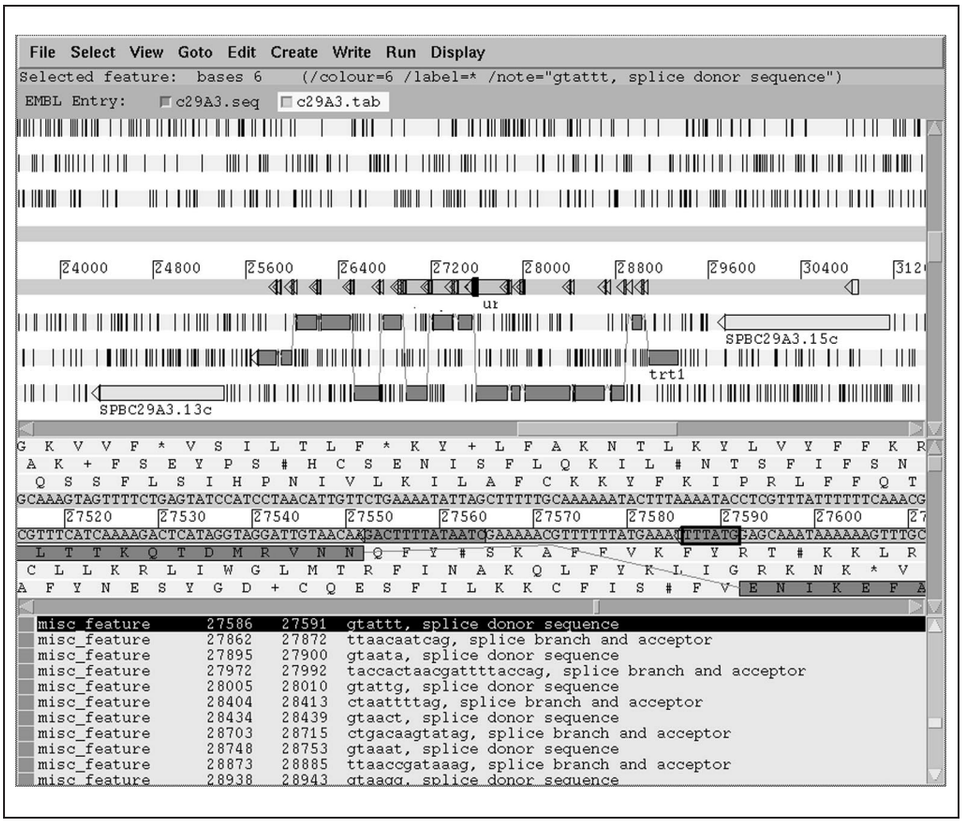

Figure 1: The main editor window of Artemis showing sequence annotation from *Schizosaccharomyces pombe*. From the top, the first panel shows a 'zoomed out' view of a DNA sequence translated into its six reading frames with vertical bars used to indicate stop codons. The view shows an overview of the highly spliced *trt1* gene. In the panel below, the DNA view shows a selected individual splice donor site in detail. At the bottom of the screen, a summary of annotation for the selected feature is shown

polycistronic transcripts (Figure 2) in the African trypanosomatid protozoan *Trypanosoma brucei* (unpublished data).

Annotation features

Artemis facilitates the intuitive manual addition of annotation to sequence data; a region of interest is highlighted and a new feature made using a 'Create' option on the menu bar. Within an editor window, annotation text can be added or altered and the type of feature can be described by assigning to it any EMBL/GenBank feature key (for instance 'CDS' or '`misc_feature`'). The large number of feature keys available allows repeats, promoters, polymorphisms and parts of genes such as untranslated regions to be handled in essentially the same manner by Artemis. Annotation is structured using EMBL/Genbank qualifiers, such as `/gene="ADH1"` and `/note="free text"`, which separate different aspects of the annotation. Annotating a diverse range of genomes requires flexibility and any number of qualifiers can be assigned to an annotated sequence feature, allowing for multiple gene names, database cross-references as well as annotator's comments. Furthermore, additional non-EMBL qualifiers can be used to increase the depth of annotation. For instance, at the Sanger Institute in-house qualifiers are used to store annotation with Gene Ontology™ terms[6] and cross-references to curated orthologues. Such extra qualifiers may be added by editing the list of available qualifiers in a configurable options file (see below). As one of its file saving options, the program will remove those extra qualifiers that are not used by EMBL/GenBank.

Feature boundaries can be altered interactively

Additional qualifiers can increase the depth of annotation

Gene prediction

Predicting genes manually from limited evidence (such as GC content) can be painstaking and is confounded further when gene models containing splicing are considered. As annotating sequence data are rarely black and white, exploring alternative feature boundaries interactively is sometimes essential. Within Artemis, the boundaries of features can be altered by dragging with the mouse pointer (if the 'direct editing' option is enabled in the Options menu of the start up window) in the DNA view. Many of the gene predictions in the *Plasmodium falciparum* genome[7] had to be refined[8,9] in this way. When trying to assess possible splice sites dragging between alternatives can quickly reveal potentially 'illegal' splice sites, as the coding sequence moves in and out of reading frame. In Figure 3, Artemis is used to manually predict the small exon from a *RIF* gene from *P. falciparum*. Small exons are hard to predict and are often missing from the output of gene prediction algorithms. In the example shown, a small exon is created, with a putative splice donor site, and merged with the second exon. In the overview the coding sequence is shown out of frame until the splice site is moved

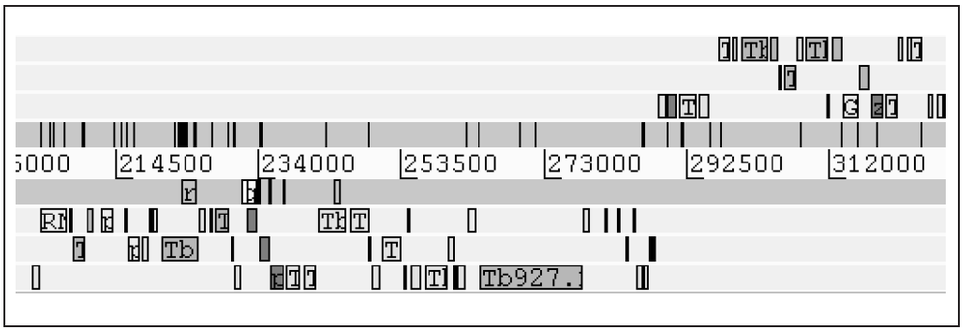

Figure 2: A 'zoomed out' overview showing the clustering of genes in *Trypanosoma brucei*. A strand switch can be seen where the coding strand switches from the reverse to forward reading frames

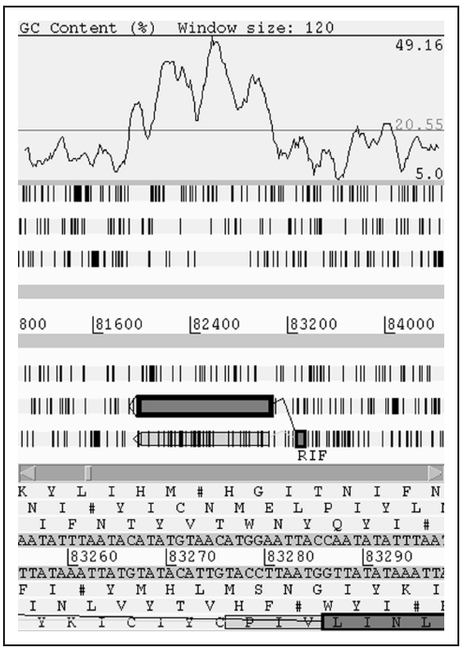

Figure 3: Editing a possible splice site in the *Plasmodium falciparum* genome. Shown highlighted is an overview of a *RIF* gene in its correct (dark) and incorrect (pale) reading frames as different splice sites are investigated by dragging with a mouse pointer in the DNA view below. Above the overview is a GC content plot, which is used to aid gene prediction. All *P. falciparum* coding sequences show higher GC content than the rest of the genome

Different layers of annotation can be viewed in parallel

in the DNA view to an alternative splice site a few bases upstream.

To arrive at a satisfactory gene model, several lines of evidence often need to be considered. For instance, the exon structures of genes in *P. falciparum* were manually edited in Artemis using proteomic data.[10] In Figure 4, the sequence of a tryptic digest fragment (LQIPSLNIIQVR) is found upstream of a predicted gene and fits a new gene model containing extra exons. Ultimately, three exons were added, to maintain the reading frame, and the intron–exon boundaries between two exons were confirmed by the sequence of the fragment. If many different types of evidence need to be used at once, such as the output from gene finding algorithms or BLAST results, they can be opened as separate entries and their annotation toggled on or off.

Figure 4 also shows how the overview can be switched to one line per entry, which allows different layers of annotation to be viewed in parallel. This is particularly useful when comparing several different evidence types such as the output from several gene finding algorithms. During the annotation of the *P. falciparum* genome, the output from three gene finding algorithms (GeneFinder (P. Green, unpublished), GlimmerM[11] and Phat[9,12]) and the results

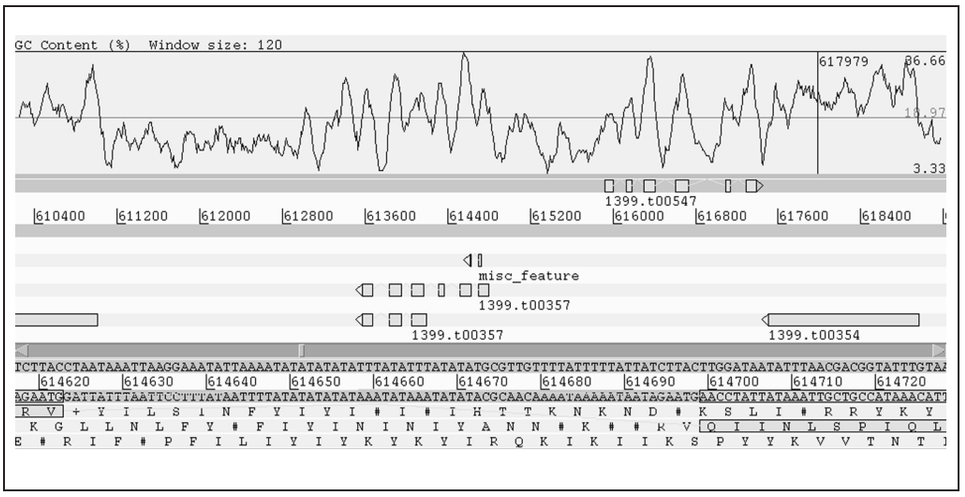

Figure 4: Use of proteomic data to refine gene models. In the overview panel, a gene model (1399.t00357) can be seen from the original annotation of *P. falciparum* in a one line per entry view. Above it, a revised version has three exons added after reviewing the gene prediction alongside the proteomic fragment (LQIPSLNIIQVR, '`misc_feature`'). The DNA view in the panel below shows the ends of two adjacent exons with the proteomic fragment shown above on the DNA sequence line. The final gene model agrees well with the GC peaks shown in the plot above

of BLAST searches (with the output from BLAST reformatted using MSPcrunch[13]) were manually reviewed to arrive at the final gene models (Figure 5). The display can easily be alternated, between showing one line per entry and all six reading frames, using a shortcut menu that is available in the DNA view and feature overview windows. For some actions, such as writing new annotation, an entry needs to be identified. In most cases, Artemis assumes that the action refers to a default entry, which can be changed by the user and is shown highlighted in yellow on the entry bar. When many entries are analysed in the one line per entry view, for clarity, it is possible to highlight in yellow the active entry within the feature overview panel.

Built-in plots can be turned on in any combination

Much of the evidence for gene predictions cannot be summarised in single lines. Artemis contains several built-in plots, including GC content, GC frame plots[14] and codon usage plots, which can be turned on in any combination. The graphs are locked to the overview so that they follow it when it is scrolled and their scale changes automatically to match that of features being viewed. Figure 6 shows how codon usage plots are used to aid coding sequence prediction in the protozoan *Leishmania major*. The codon usage plots contain a colour-coded graph for each reading frame; thus predictions can be made based on finding open reading frames which have codon usage peaks in the correct frame. In the *P. falciparum* genome project, percentage GC plots were essentially the only tool available for predicting exons that were not found using Genefinder, Phat or GlimmerM. The various graphs in Artemis can also be used to identify sequences that appear unusual relative to the rest of a genome. In this way, evidence of large-scale transposition or laterally acquired sequences can be visualised. For instance, in the *Salmonella typhi* genome,[15] GC content, GC deviation and Karlin Signature difference[16] plots can all be used to visualise pathogenicity islands, containing virulence-determinants, such as the previously described SPI2[17] (Figure 7).

Customising the display

The amount of screen space used can be changed by adding or removing the DNA view, the Feature Overview, the Feature

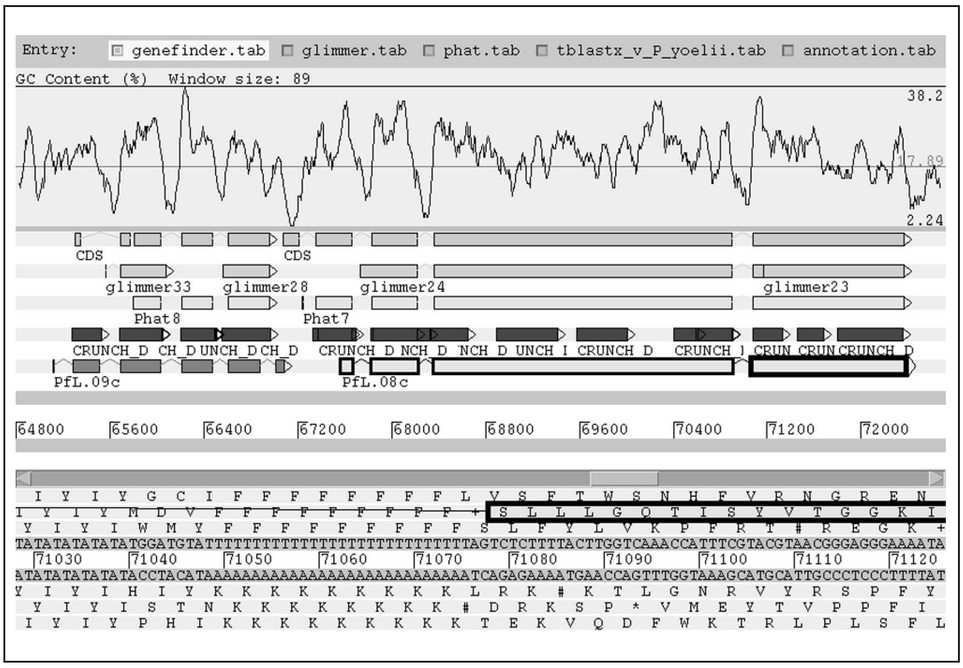

Figure 5. Several lines of evidence are compared to aid gene prediction. The overview window shows the output from Genefinder, GlimmerM, Phat and BLAST being used to predict two genes (PfL.09c and PfL.08c). Each line of evidence is in a separate entry and can be toggled on or off by clicking on the entry bar at the top of the editor window

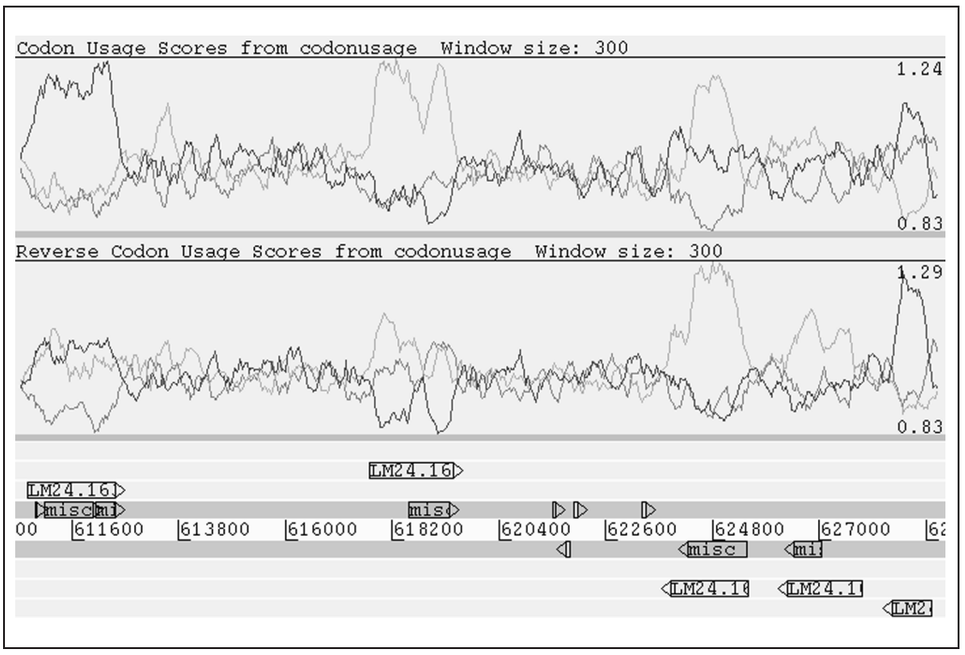

Figure 6: Prediction of genes in *Leishmania major* using codon usage plots. The overview panel shows five genes in their respective reading frames that were predicted manually based on their codon usage plots. The peaks refer to specific reading frames, highlighting open reading frames with high coding potential

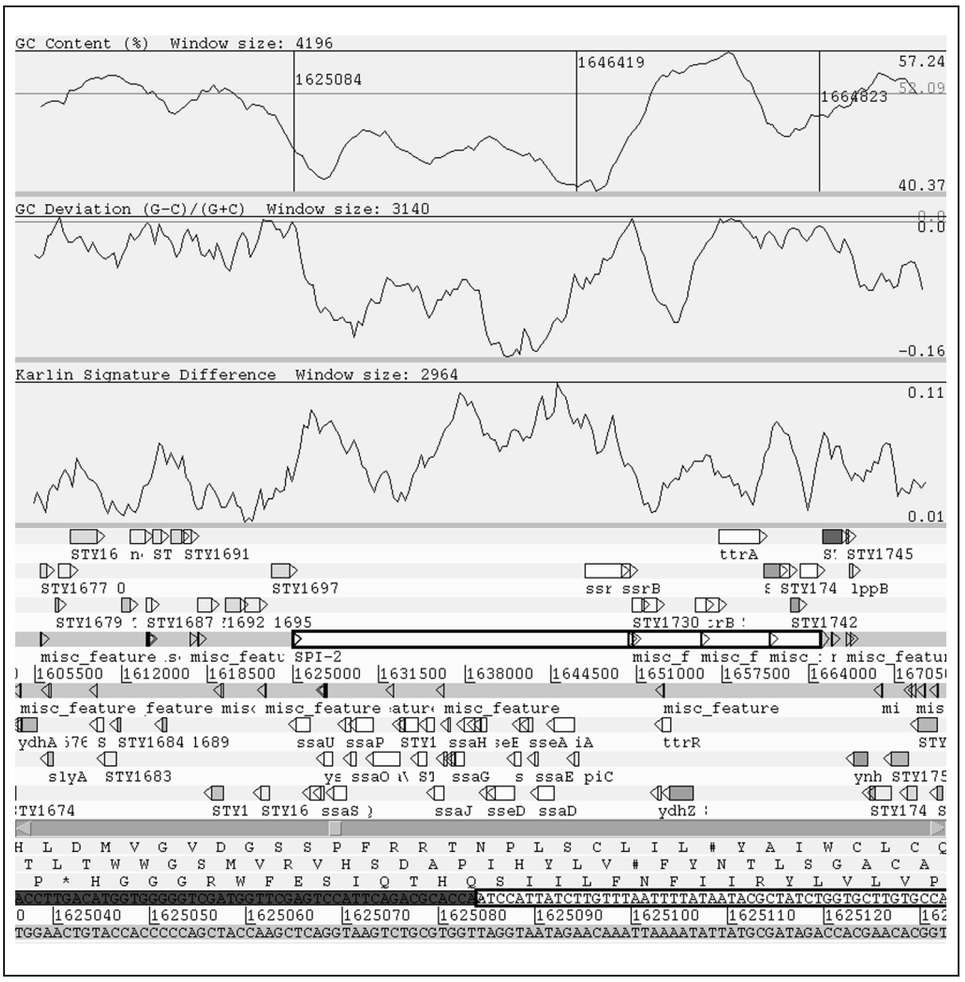

Figure 7: Changes in the GC content, GC deviation and the Karlin Signature difference reveal the location of a pathogenicity island in the *Salmonella typhi* genome. A white horizontal bar along the DNA line (grey) shows the exact position of the SPI-2 island within the annotation

List or various graphical analyses (above). For clarity, features in complex annotation can be distinguished by applying colours. Artemis contains a default colour scheme with 18 preset colours. Within a session, changes can be made globally or individually to the '`/colour=`' features by entering RGB colour codes or by using one of the 18 preset codes. Alternatively, there is an options file (below) where default values can be changed, particularly useful if the details of a feature are obscured.[18]

A navigator window allows the user to jump along a sequence

Options file

Upon start-up, Artemis looks in up to five places for files containing configuration information. One of these, and perhaps the easiest to make changes to, is a file called `options`, `options.txt` or `options.text`, which on PC or Macintosh computers is found in the directory within which Artemis is installed. On machines using Unix, Artemis will look in the directory from where it was launched. In addition to changing colours, there are many other configurable options available, for instance the list of feature keys or qualifiers can be edited to include ones not used by EMBL or GenBank.

GETTING MORE FROM ARTEMIS

Editing sequences

Unfortunately, annotated sequences do change; base ambiguities often get resolved or perhaps mistakes are uncovered. Bases can be deleted or inserted, either directly or by inserting them from a text file. When this is done, care should of course be taken as changes to a coding sequence could invalidate previous descriptions of a gene's function. For large sequences, it is often convenient to split them into more manageable sections. This can also be done using the Edit menu of Artemis. A region of bases is selected and a copy made that opens automatically in a new Artemis Edit window. The subsequence includes any annotated features that are located within the original selection. Features that overlap the ends of a selected range are truncated accordingly.

Large sequences can be split into more manageable subsequences

Analysing and navigating around sequences

As well as using the scroll bars to move along a sequence, the user can use the items in the Goto menu. A navigator window can be opened, which allows the user to jump along a sequence, automatically scrolling all of the views to centre at a particular location based on one the following: a specific base coordinate, the next feature that contains an entered string in its annotation, the next feature that has a specified feature key such as `tRNA` or `CDS`, or the next occurrence of a specified base or amino acid pattern.

To perform further analyses, annotation or for writing out sequences into files it is necessary to be able to select sequences based on a range of criteria. Most simply, bases are highlighted with a mouse or all the bases of a sequence may be highlighted using the Select menu. Individual features can also be selected; for instance all CDS features can be selected at once. More elaborate 'querying' is possible using the Feature Selector function, which allows searches based on feature types, size cut-offs or text strings within annotation to be combined. It is also possible to select overlapping features; such as those that represent the intersection of two sets of analyses. In this way, for example, if an entry containing annotated coding sequences and an entry containing single nucleotide polymorphisms (SNPs), are viewed, those SNPs found in coding sequences can be easily selected (Figure 8a). Accordingly, it is possible to select those CDS in which SNPs have been identified (Figure 8b). The selected sequences can then be written out to a file for further analyses.

UNIX VERSION

The Unix version of Artemis is slightly different from the versions for PC or Macintosh computers. As well as being

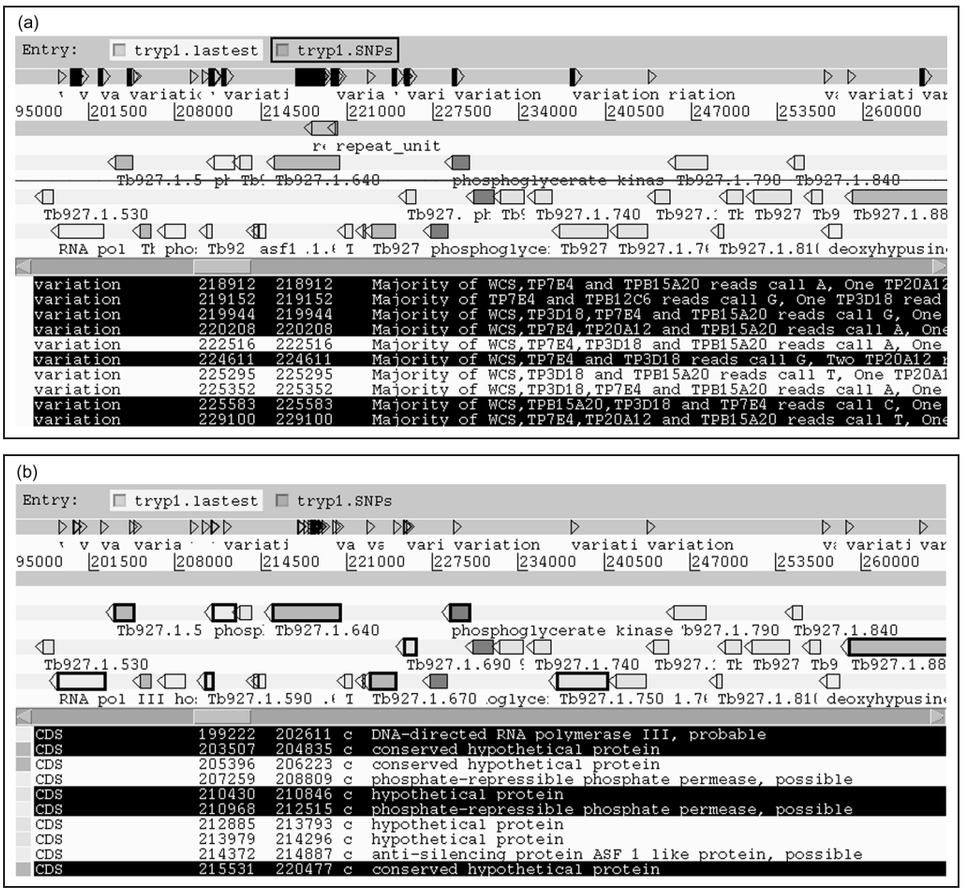

Figure 8: Selecting overlapping data sets. Artemis can be used to select the intersection between a CDS data set, which is shown as boxes drawn on the three reverse reading frames and SNP data, which are shown on the forward DNA line above the reading frames. Note that many of the SNPs cannot be resolved as individual features until the view is zoomed in further. In (a) annotation for *Trypanosoma brucei* is opened first and all coding sequences selected. A second file is then opened and overlapping features selected, ie only those SNPs that are found in coding sequences. In (b) coding sequences that contain SNPs are selected by following a similar procedure but loading the SNP data first

able to specify more options on the command line when the program is launched, it also contains an addition: the Run menu. This allows programs such as BLAST or FASTA to be launched from within Artemis. The results of these searches can then be viewed from within Artemis on all platforms.

CONCLUSION

The use of Artemis as a sequence viewer as well as an annotation tool is widespread. Because it is a stand-alone package, anyone can quickly get to grips with using it to get a very detailed view of sequence data and the features that have been annotated on it. Artemis places particular emphasis on being able to directly view the DNA sequence at all times in fine detail, making the manual prediction of genes more effective and helping the user to spot unusual DNA features. Gene models can be edited interactively in the context of supporting evidence, allowing the discovery of many genes or exons that have been missed by gene prediction algorithms. Furthermore, large sections of DNA can be viewed and

analysed at once, allowing unusual features resulting from DNA rearrangements or lateral acquisition events to be identified.

The flexibility of Artemis has allowed it to be applied to a large number of genome projects and the software is continually evolving based on feedback from the annotators and other users. Moreover, its portability – a driving principle behind the program's development – has helped hundreds of users, many laboratory-based, to have the same access to genome projects as their colleagues might have at genome centres.

Acknowledgments
The development of Artemis is funded by the Wellcome Trust.

References

1. URL: http://www3.ncbi.nlm.nih.gov/Entrez/Genome/org.html
2. Rutherford, K., Parkhill, J., Crook, J. *et al.* (2000), 'Artemis: Sequence visualization and annotation', *Bioinformatics*, Vol. 16(10), pp. 944–945.
3. URL: http://www.sanger.ac.uk/Software/Artemis
4. URL: http://www.gnu.org/copyleft/gpl.html
5. Wood, V., Gwilliam, R., Rajandream, M. A. *et al.* (2002), 'The genome sequence of *Schizosaccharomyces pombe*', *Nature*, Vol. 415(6874), pp. 871–880.
6. Ashburner, M., Ball, C. A., Blake, J. A. *et al.* (2000), 'Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium', *Nat. Genet.*, Vol. 25(1), pp. 25–29.
7. Gardner, M. J., Hall, N., Fung, E. *et al.* (2002), 'Genome sequence of the human malaria parasite *Plasmodium falciparum*', *Nature*, Vol. 419(6906), pp. 498–511.
8. Hall, N., Pain, A., Berriman, M. *et al.* (2002), 'Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13', *Nature*, Vol. 419(6906), pp. 527–531.
9. Hyman, R. W., Fung, E., Conway, A. *et al.* (2002), 'Sequence of *Plasmodium falciparum* chromosome 12', *Nature*, Vol. 419(6906), pp. 534–537.
10. Lasonder, E., Ishihama, Y., Andersen, J. S. *et al.* (2002), 'Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry', *Nature*, Vol. 419(6906), pp. 537–542.
11. Salzberg, S. L., Pertea, M., Delcher, A. L. *et al.* (1999), 'Interpolated Markov models for eukaryotic gene finding', *Genomics*, Vol. 59(1), pp. 24–31.
12. Cawley, S. E., Wirth, A. I. and Speed, T. P. (2001), 'Phat – a gene finding program for *Plasmodium falciparum*', *Mol. Biochem. Parasitol.*, Vol. 118(2), pp. 167–174.
13. Sonnhammer, E. L. and Durbin, R. (1994), 'A workbench for large-scale sequence homology analysis', *Comput. Appl. Biosci.*, Vol. 10(3), pp. 301–307.
14. 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*, Vol. 30(1–3), pp. 157–166.
15. Parkhill, J., Dougan, G., James, K. D. *et al.* (2001), 'Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18', *Nature*, Vol. 413(6858), pp. 848–852.
16. Karlin, S. and Burge, C. (1995), 'Dinucleotide relative abundance extremes: A genomic signature', *Trends Genet.*, Vol. 11(7), pp. 283–290.
17. Shea, J. E., Hensel, M., Gleeson, C. and Holden, D. W. (1996), 'Identification of a virulence locus encoding a second type III secretion system in *Salmonella typhimurium*', *Proc. Natl Acad. Sci. USA*, Vol. 93(6), pp. 2593–2597.
18. Mural, R. J. (2000), 'ARTEMIS: A tool for displaying and annotating DNA sequence.' *Brief. Bioinform.*, Vol. 1(2), pp. 199–200.