BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 23 2008, pages 2672–2676
doi:10.1093/bioinformatics/btn529

Genome analysis

Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database

Tim Carver*, Matthew Berriman, Adrian Tivey, Chinmay Patel, Ulrike Böhme, Barclay G. Barrell, Julian Parkhill and Marie-Adèle Rajandream

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Received on August 10, 2008; revised on September 19, 2008; accepted on October 6, 2008

Advance Access publication October 9, 2008

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Artemis and Artemis Comparison Tool (ACT) have become mainstream tools for viewing and annotating sequence data, particularly for microbial genomes. Since its first release, Artemis has been continuously developed and supported with additional functionality for editing and analysing sequences based on feedback from an active user community of laboratory biologists and professional annotators. Nevertheless, its utility has been somewhat restricted by its limitation to reading and writing from flat files. Therefore, a new version of Artemis has been developed, which reads from and writes to a relational database schema, and allows users to annotate more complex, often large and fragmented, genome sequences.

Results: Artemis and ACT have now been extended to read and write directly to the Generic Model Organism Database (GMOD, http://www.gmod.org) Chado relational database schema. In addition, a Gene Builder tool has been developed to provide structured forms and tables to edit coordinates of gene models and edit functional annotation, based on standard ontologies, controlled vocabularies and free text.

Availability: Artemis and ACT are freely available (under a GPL licence) for download (for MacOSX, UNIX and Windows) at the Wellcome Trust Sanger Institute web sites:
http://www.sanger.ac.uk/Software/Artemis/
http://www.sanger.ac.uk/Software/ACT/

Contact: artemis@sanger.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Artemis (Berriman and Rutherford, 2003; Rutherford *et al*., 2000) and the Artemis Comparison Tool (Carver *et al*., 2005) (ACT) have gathered increasing number of users since their initial release with more than 33 000 downloads in 2007 (Fig. 1) from over 90 countries. Artemis is a sequence viewer and annotation tool. It is used primarily not only for genome annotation but also for DNA visualization and analysis, and as a teaching aid. The user can view the sequence and features in context at different levels of granularity from the amino acids to the complete genome level, with continuous zooming between these levels. Artemis includes clear and interactive navigation, customizable search facilities, predefined and user-configurable graphical plots and statistical overviews. ACT is an extension that makes use of the Artemis components to display pairwise comparisons between two or more sequences. Therefore, ACT inherits a lot of the functionality found in Artemis. ACT is used to identify and analyse regions of similarity and difference between genomes and to explore conservation of synteny, in the context of the entire sequences and their annotation.

Artemis and ACT were originally designed to read and write flat files in the common sequence formats (EMBL, GenBank, FASTA and GFF3) to display and edit sequence and annotation, and they continue to be able to work in this mode. This works well for many uses but does restrict Artemis to single user access when annotating sequences. A previous way around this problem was to split the files up and merge them back after editing. This is far from ideal, particularly for large multi-user projects. To maximize the potential of Artemis as a tool for community annotation, Artemis has been extended to connect to a relational database, and now supports reading and writing to the GMOD Chado schema.

Artemis and ACT use the iBatis Data Mapper framework (http://ibatis.apache.org) to map Java objects to SQL parameters and result sets. For each of the tables in the Chado schema that are used, an XML descriptor file is used to define the mapping. This has the advantage of simplifying the Artemis code by separating it from the SQL and it also means that there is complete control over the SQL queries, without modification of the Artemis code.

An alpha-version of the software presented here was robustly tested during a week-long *Plasmodium falciparum* (malaria) re-annotation workshop in October 2007. Fixes, additional functionality and further optimizations stemming from this were then included in the new releases of Artemis v.10 and ACT v.7.

2 IMPLEMENTATION

2.1 Artemis and ACT reading and writing in database mode

Artemis and ACT can be launched in database mode from a web launch or from the command line. Connection to the database is established via Java Database Connectivity (JDBC) and by specifying the database location and the JDBC driver, for example:

```
art -Dchado="hostname:port/databaseName?userName" \
  -Djdbc.drivers=org.postgresql.Driver -Dibatis
```

*To whom correspondence should be addressed.

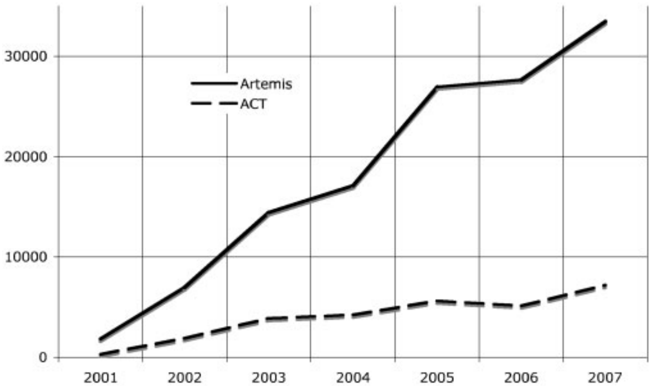


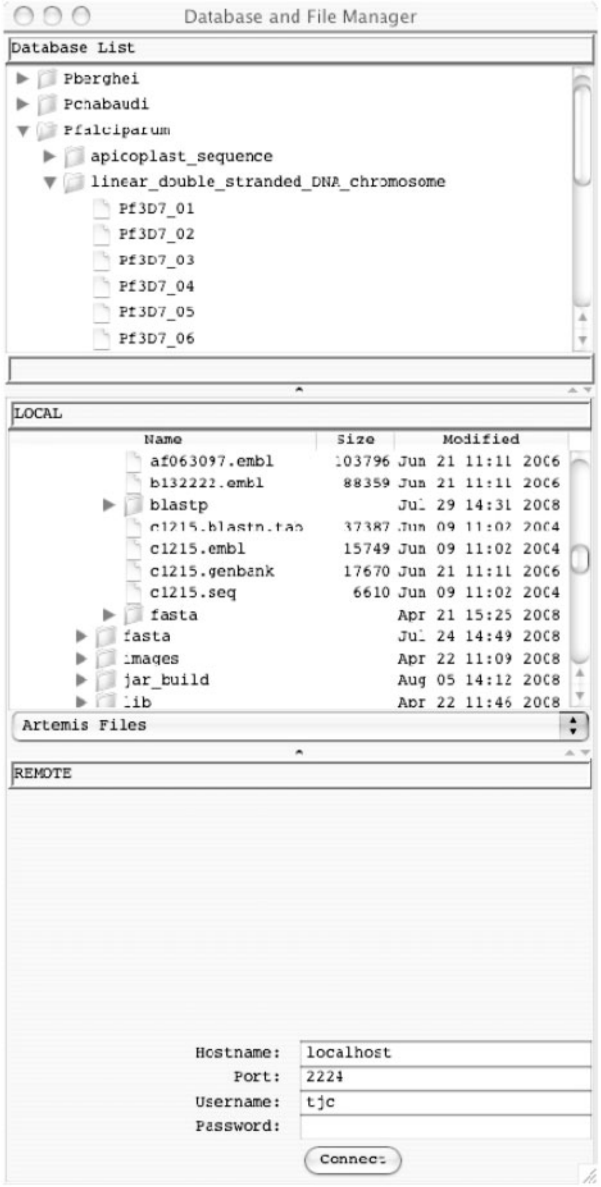
Fig. 1. Download statistics for Artemis and ACT for each year from 2001–2007.

This prompts for a password for the given username. For a read-only database the flag –Dread_only will suppress the password prompt. On successfully logging in a database and file manager window opens showing an expandable organism tree (Fig. 2). Double clicking on the sequence names opens them in Artemis. ACT currently still works with comparison flat files but the sequences themselves can be dragged from the database manager into the ACT file requestor. It is possible to store comparative regions in Chado so that ACT could in the future read this information from the database as well. All of the sequence and annotations are read from the database when the interface is opened. The exceptions to this are the cluster and orthologue and paralogue data that are read lazily; that is that they are only fully read in from the database when they are viewed in the Gene Builder. Loading these data into Artemis in this way speeds up the initial process of opening the sequence in the interface. For future releases of this software, lazy loading of the majority of the annotation will be explored. This is likely to become more important as the database grows.

Annotation within Artemis traditionally centres around 'feature keys' used by the International Nucleotide Sequence Database Collaboration (INSDC), which label-specific annotation types within a flat-file format. In particular, protein-coding sequences or 'CDS' were the default annotation objects, although other feature keys could be specified. The Chado schema is designed to store descriptions of any feature that can be located in a genome. For instance, it holds a complete gene model consisting of the gene itself, transcripts, exons, untranslated regions (UTRs) and polypeptides. It makes use of the Sequence Ontology (SO, http://www.sequenceontology.org/) (Eilbeck *et al.*, 2005) to represent these data, such that a full hierarchy of genome features can be added. The CDS features themselves are not stored, as they can then be inferred from the coordinates of the other features. SO is designed to describe biological sequences including located sequence features, for example, polymorphisms, polypeptide domains and predictions. All these feature types can therefore be stored in Chado and viewed in Artemis and ACT.

In database mode, the majority of the functionality and appearance of Artemis has been maintained (Fig. 3). However, Artemis has been changed to be able to handle the hierarchy of constituent features relating to genes (exons, transcripts, polypeptides, etc.). It was also a requirement that it was able to present the annotation in a structured manner and able to handle the main underlying ontologies, i.e. gene, sequence and relationship.

A Gene Builder tool has been developed for displaying and manipulating the gene hierarchy and associated annotation. A gene is created by highlighting a base range and selecting, from the 'Create' menu, the 'Gene Model From Base Range' option. The basic constituent features are created; i.e. gene, transcript, CDS and polypeptide. Artemis joins exon features from the underlying database that are part of the same transcript and represents them as a CDS feature, which is shown on the reading frame lines in its main display. For clarity, the polypeptide and transcript features are hidden in the feature display window and appear greyed out in the

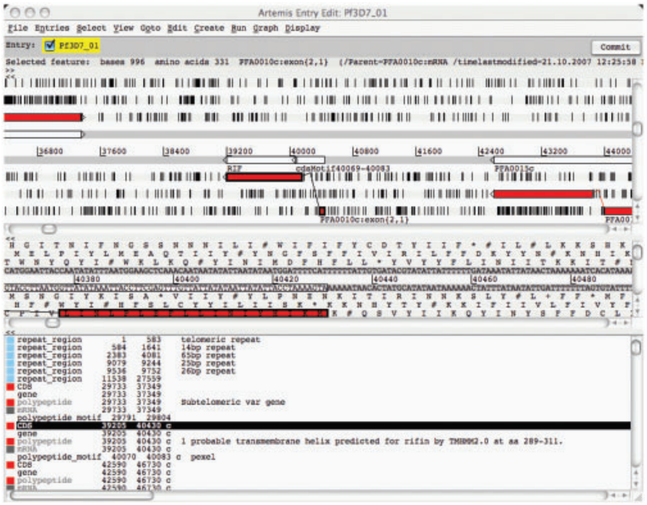


```
▶ Pberghei
▶ Pchabaudi
▼ Pfalciparum
    ▶ apicoplast_sequence
    ▼ linear_double_stranded_DNA_chromosome
        Pf3D7_01
        Pf3D7_02
        Pf3D7_03
        Pf3D7_04
        Pf3D7_05
        Pf3D7_06
```

Name	Size	Modified
af063097.embl	103796	Jun 21 11:11 2006
b132222.embl	88359	Jun 21 11:11 2006
▶ blastp		Jul 29 14:31 2008
c1215.blastn.tab	37387	Jun 09 11:02 2004
c1215.embl	15749	Jun 09 11:02 2004
c1215.genbank	17670	Jun 21 11:11 2006
c1215.seq	6610	Jun 09 11:02 2004
▶ fasta		Apr 21 15:25 2008
▶ fasta		Jul 24 14:49 2008
▶ images		Apr 22 11:09 2008
▶ jar_build		Aug 05 14:12 2008
▶ lib		Apr 22 11:46 2008

Hostname: localhost
Port: 2224
Username: tjc
Password:

Fig. 2. The Artemis database and file manager. The organisms in the database are shown at the top. Double clicking on the chromosome opens it up in Artemis. The file management system also provides access to local (middle panel) and remote file systems (via ssh; bottom panel).



Entry: ☑ Pf3D7_01

```
Selected feature:  bases 996  amino acids 331  PFA0010c:exon(2,1)  (/Parent=PFA0010c:mRNA /timelastmodified=21.10.2007 12:25:58
```

```
repeat_region          1      583    telomeric repeat
repeat_region        584     1641    14bp repeat
repeat_region       2383     4081    65bp repeat
repeat_region       9079     9244    25bp repeat
repeat_region       9536     9752    26bp repeat
repeat_region      11538    27559
CDS                29733    37349
gene               29733    37349
polypeptide        29733    37349    Subtelomeric var gene
mRNA               29733    37349
polypeptide_motif  29791    29804
CDS                39205    40430 c
gene               39205    40430 c
polypeptide        39205    40430 c  1 probable transmembrane helix predicted for rifin by TMHMM2.0 at aa 289-311.
mRNA               39205    40430 c
polypeptide_motif  40070    40083 c  pexel
CDS                42590    46730 c
gene               42590    46730 c
polypeptide        42590    46730 c
mRNA               42590    46730 c
```

Fig. 3. Snapshot of Artemis in database mode displaying chromosome 1 of *P. falciparum*. The sequence and feature displays are identical to the standard Artemis. There is a 'Commit' button at the top right of the interface for writing back to the database. The top feature display panel shows a zoomed-out representation of the region. Below this is a zoomed-in view showing the nucleotide level with the six frames of translation. The lower panel shows a scrollable list of the features. These panels are all linked, so that if a feature is selected by double clicking on it the other windows scroll to the feature. The polypeptide and transcript are hidden from the feature display but appear (greyed out) in the feature list at the bottom.

feature list as illustrated in Figure 3. Note there is an option in the feature display pop-up menu to control what features are hidden. This can be used to hide features that are not of interest during a session.

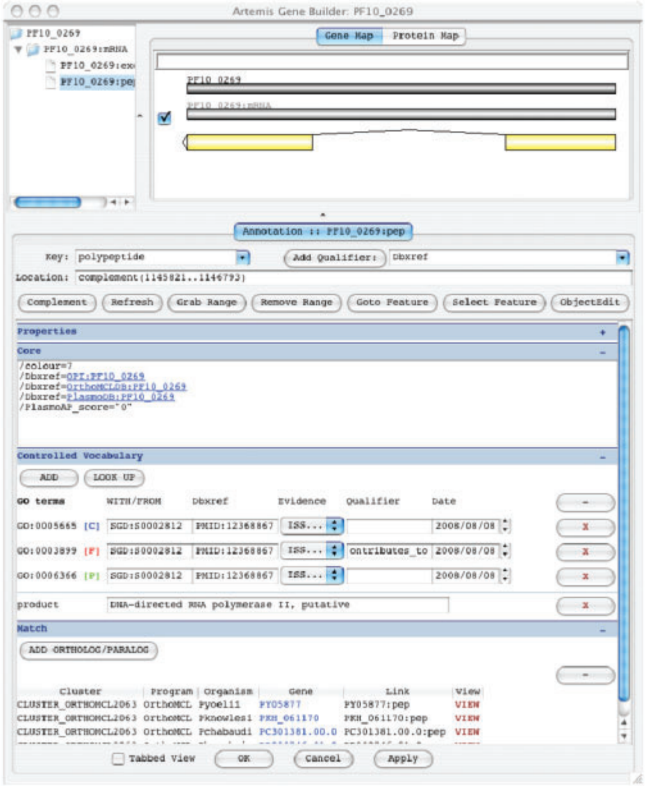
2.2 Artemis Gene Builder

A Gene Builder window for a selected gene feature can be opened from the 'Edit' menu by selecting the 'Selected Feature in Editor' option. The gene structure and feature annotation are two distinct parts of the Gene Builder (Fig. 4), and are described below. In general the majority of the annotation is attached to the polypeptide, although the system allows annotation to any object within the gene hierarchy. Therefore when a CDS is selected, and the Gene Builder opened for the corresponding gene, it displays the annotation on the polypeptide.

2.2.1 *Gene hierarchy and structure*

In the top left-hand side of the Gene Builder is a tree structure of the gene model. To the right of this is a Gene Map, a graphical representation of the features. These show the gene hierarchy as described by SO. A feature can be selected from either the tree or the graphical view. The annotation for the selected feature is displayed in the bottom part of the Gene Builder.

Another new feature the Gene Builder brings to Artemis is the ability to display protein maps (Fig. 5). If there is any polypeptide predictions associated with a feature then a Protein Map tab is shown. When selected this displays the information about the protein and the predictions, e.g. InterPro (Mulder *et al*., 2007), SignalP (Bendtsen *et al*., 2004) and others.



Key: polypeptide

Add Qualifier: Dbxref

Location: complement(1145821..1146793)

```
/colour=7
/Dbxref=OPI:PF10_0269
/Dbxref=OrthoMCLDB:PF10_0269
/Dbxref=PlasmoDB:PF10_0269
/PlasmoAP_score="0"
```

GO terms	WITH/FROM	Dbxref	Evidence	Qualifier	Date
GO:0005665 [C]	SGD:S0002812	PMID:12368867	ISS...		2008/08/08
GO:0003899 [F]	SGD:S0002812	PMID:12368867	ISS...	ontributes_to	2008/08/08
GO:0006366 [P]	SGD:S0002812	PMID:12368867	ISS...		2008/08/08

product: DNA-directed RNA polymerase II, putative

Cluster	Program	Organism	Gene	Link	View
CLUSTER_ORTHOMCL2063	OrthoMCL	Pyoelii	PY05877	PY05877:pep	VIEW
CLUSTER_ORTHOMCL2063	OrthoMCL	Pknowlesi	PKH_061170	PKH_061170:pep	VIEW
CLUSTER_ORTHOMCL2063	OrthoMCL	Pchabaudi	PC301381.00.0	PC301381.00.0:pep	VIEW

□ Tabbed View

Fig. 4. The Gene Builder showing a gene hierarchy at the top and underneath the annotation for the associated polypeptide.

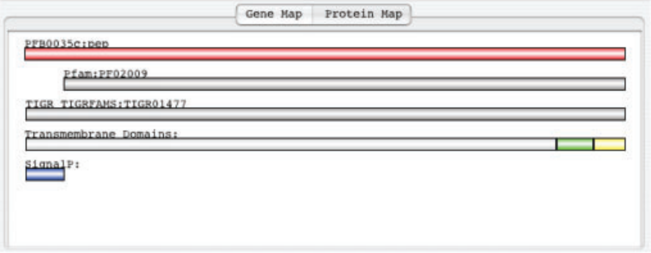


Fig. 5. An example of a Protein Map in the Gene Builder. The details of the domain predictions are shown in mouse-over tool-tips. The protein database features (e.g. Pfam, Prosite) can be clicked on to open that entry in a browser window.

Structural changes can be carried out in the graphical view. Feature ends can be selected and dragged to adjust their coordinates.

On right clicking on this area a pop-up menu allows features to be added or deleted in the gene model. Exons are added by highlighting a range in the Gene Builder window, clicking the left mouse button and dragging, and then adding an exon from the pop-up menu. The exon coordinates can be refined by dragging the exon boundary in the zoomed-in (base-level) view of the main Artemis window. Multiple transcripts (Fig. 6) and their associated exons can be added to the gene model. In previous versions of Artemis, it could prove difficult to build and annotate overlapping splice forms from a single gene. Therefore, in order to make it easier to deal with overlapping

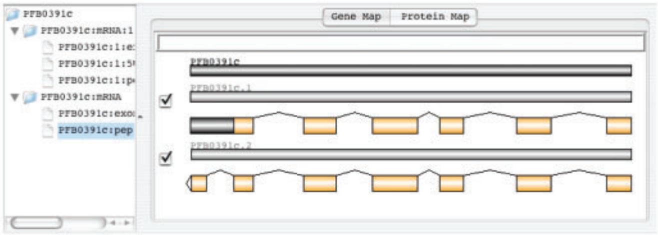
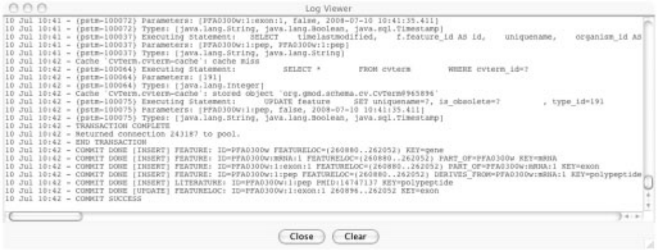


Fig. 6. Example of a gene with multiple transcripts in the graphical view of the gene hierarchy in the Gene Builder. Each transcript is a child node of the gene in the tree on the left with its associated features displayed when expanded. To assist in building multiple transcripts the checkboxes to the left of each transcript are used to hide and show them in the main Artemis window.



```
10 Jul 10:41 - (pstm-100072) Parameters: [PFA0300w:1:exon:1, false, 2008-07-10 10:41:35.411]
10 Jul 10:41 - (pstm-100072) Types: [java.lang.String, java.lang.Boolean, java.sql.Timestamp]
10 Jul 10:41 - (pstm-100037) Executing Statement:   SELECT    timelastmodified,    f.feature_id AS id,    uniquename,    organism_id AS
10 Jul 10:41 - (pstm-100037) Parameters: [PFA0300w:1:pep, PFA0300w:1:pep]
10 Jul 10:41 - (pstm-100037) Types: [java.lang.String, java.lang.String]
10 Jul 10:42 - Cache 'CvTerm.cvterm-cache': cache miss
10 Jul 10:42 - (pstm-100064) Executing Statement:        SELECT *       FROM cvterm       WHERE cvterm_id=?
10 Jul 10:42 - (pstm-100064) Parameters: [191]
10 Jul 10:42 - (pstm-100064) Types: [java.lang.Integer]
10 Jul 10:42 - Cache 'CvTerm.cvterm-cache': stored object 'org.gmod.schema.cv.CvTerm@965896'
10 Jul 10:42 - (pstm-100075) Executing Statement:     UPDATE feature     SET uniquename=?, is_obsolete=?        , type_id=191
10 Jul 10:42 - (pstm-100075) Parameters: [PFA0300w:1:pep, false, 2008-07-10 10:41:35.411]
10 Jul 10:42 - (pstm-100075) Types: [java.lang.String, java.lang.Boolean, java.sql.Timestamp]
10 Jul 10:42 - TRANSACTION COMPLETE
10 Jul 10:42 - Returned connection 243187 to pool.
10 Jul 10:42 - END TRANSACTION
10 Jul 10:42 - COMMIT DONE [INSERT] FEATURE: ID=PFA0300w FEATURELOC=(260880..262052) KEY=gene
10 Jul 10:42 - COMMIT DONE [INSERT] FEATURE: ID=PFA0300w:mRNA:1 FEATURELOC=(260880..262052) PART_OF=PFA0300w KEY=mRNA
10 Jul 10:42 - COMMIT DONE [INSERT] FEATURE: ID=PFA0300w:1:exon:1 FEATURELOC=(260880..262052) PART_OF=PFA0300w:mRNA:1 KEY=exon
10 Jul 10:42 - COMMIT DONE [INSERT] FEATURE: ID=PFA0300w:1:pep FEATURELOC=(260880..262052) DERIVES_FROM=PFA0300w:mRNA:1 KEY=polypeptide
10 Jul 10:42 - COMMIT DONE [INSERT] LITERATURE: ID=PFA0300w:1:pep PMID:14747137 KEY=polypeptide
10 Jul 10:42 - COMMIT DONE [UPDATE] FEATURELOC: ID=PFA0300w:1:exon:1 260896..262052 KEY=exon
10 Jul 10:42 - COMMIT SUCCESS
```

Fig. 7. Snapshot of the Artemis log viewer, using log4j to generate log messages and errors. It records SQL statements and can be used to monitor changes to the database. The last seven lines in this log are generated to show the changes made to the database in a more human readable format.

features in the Gene Builder, a checkbox for each transcript can be used to show and hide them in the main Artemis window.

2.2.2 Location, keys and annotation For the selected feature in the gene model, the location, key and annotation are displayed under the gene model hierarchy and graphical structure. To emphasize which of the features the annotation shown belongs to, its name is displayed above the feature key selector. The coordinates and feature key can be changed here.

Instead of the free text used in the original Artemis editor, the Gene Builder provides a structured means of annotating. There are four annotation sections in the Gene Builder described below. A check box at the bottom of the Gene Builder changes between a scrollable and a tabbed view of the sections. In the scrollable view, the sections have hide (−) and show (+) buttons and empty sections are automatically hidden.

(1) Properties: this contains feature properties, such as synonyms and time last modified. Synonyms are categorized using the controlled vocabulary tables in Chado. The time last modified is updated when a change to that feature is written back to the database. It is often desirable for data from previous annotation versions to still be available, therefore in this section of the Gene Builder there is a checkbox to mark features and entire gene models as obsolete. This hides them from view in the main Artemis window, but they are still available from the feature list at the bottom.

(2) Core: the core annotation contains any annotation that does not fit into the other sections. For example, free-text comments and cross-links to the scientific literature. Hyperlinks are provided for databases (e.g. UniProt, EMBL, PubMed and others defined in the options file) that open up a local browser.

(3) Controlled vocabulary (CV): the CV module in the schema is concerned with controlled vocabularies or ontologies. Artemis uses biological ontologies to allow very precise and expressive annotation. A form is provided for adding and deleting Gene Ontology (GO; http://www.geneontology.org/) descriptions, product, Riley class and other controlled annotations, which are stored in the CV tables. When adding a term to a feature the CV (e.g. product) and a keyword (e.g. histone) are prompted for. The term to be added is then selected from a drop down list of terms containing the word or phrase. Additional controlled curations can be shown if their CV name in Chado is prefixed with 'CC_' (e.g. CC_workshop).

(4) Match: orthologue and paralogue links can be added to other genes in the database in this section. Externally calculated paralagous gene clusters can also be displayed here. There are links for opening the Gene Builder for each matched gene entry or for opening a separate Artemis instance displaying the gene and the surrounding features. Future developments to the Gene Builder will add to this and incorporate annotation transfer mechanisms. For example, the user will be able to specify the annotation to transfer to or from the selected genes in the list of matches.

2.3 Writing to the database

When a feature or qualifier is changed, added or deleted the 'Commit' button (on the top tool bar) changes colour to red. Changes only get written back to the database when this button is clicked. There is also an option under the 'File' menu to 'Commit To Database'. In ACT, there is no commit button and committing back to the database is carried out from the menus.

If there is an error during the commit then an option to force commit is provided and Artemis will commit what it can. The Apache Log4j (http://logging.apache.org/log4j) framework has been used to provide logging information in the Artemis log window (Fig. 7). Using the standard configuration file (log4j.properties) the logging can be directed to a file. The database read and write statements are logged as SQL as well as a more human readable notation of the write (insert, update and delete) statements. This is useful for monitoring progress and generating statistics for the annotation process.

In Artemis and ACT, additional features from files can be read in and overlaid on the database entry. If they are of the correct sequence ontology they can then be bulk loaded into the database. This is done in the interface by selecting the features and moving them to the database entry. These features can then be committed to the database.

Flat files are still needed for submission and to run external analyses on. It is therefore possible to write out EMBL flat files from the database entry in Artemis. The default option is to flatten the gene hierarchy and transfer the annotation onto a CDS feature.

2.4 Community annotation

Multiple users can launch Artemis and ACT clients and both query and modify the database simultaneously. This enables multiple annotators to work on the same sequence at the same time. This has been stress tested and used in a *Plasmodium falciparum* (malaria) genome re-annotation workshop (October 2007). More than 30 scientists used Artemis clients to simultaneously connect to a single database and annotate the genome. Access to the database was also made available to the course participants *via* a virtual private network (VPN) after the course. External public access to this data is also possible (read only) via local instances of Artemis connecting to a copy of the database at the Sanger Institute (http://www.sanger.ac.uk/Software/Artemis/databases/).

For selected features, BLAST and FASTA results can be generated in Artemis via the Run menu. The results are written to a file. For community projects, these need to be stored in a central file system or copied to the different sites involved. A future development for Artemis is likely to make use of optionally storing search results files directly within the database, or a specific results file server to simplify data management within large shared projects.

To ensure that users do not overwrite each other's annotation, Artemis records the time each feature was last modified. This is then checked against the database time stamp of the features before the changes are written to the database. If the corresponding feature in the database has been changed by another user, Artemis will warn that this has occurred and ask whether to continue with the commit process, thus allowing a forced overwrite of previous changes, where necessary.

3 DISCUSSION

From their conception Artemis and ACT have been designed to be extensible. This has enabled them to have functionality added as needed and to evolve with the requirements of various levels of users. Artemis (version 10) and ACT (version 7), using iBatis mappings and JDBC, can now connect to and read and write to a Chado relational database. This allows access from multiple clients to sequence data. A form of optimistic versioning is used so that Artemis will warn a user if another user has modified the feature they are updating during that Artemis session.

Annotation is displayed and edited in a structured manner in the new Artemis Gene Builder using predefined and standard ontologies. Links within the database can be used to open up separate Gene Builders and Artemis windows for individual regions. Hyperlinks to external databases launching a local browser window allow rapid and easy linking to relevant external data sources.

Using the gene hierarchy allows more fine-grained annotation, with multiple alternate transcripts, CDSs, UTRs, etc. In addition, the standardized ontology-based schema allows Artemis to be both a powerfully expressive and extremely precise annotation tool.

Artemis and ACT can be used to display and annotate both eukaryotes and prokaryotes. A eukaryotic example has been presented here but, because SO terms define the data model, Chado can be used to describe prokaryotes as well. Prokaryote genes are created in the same way and represented with a simple gene model with a gene, transcript and a single CDS. Prokaryote annotations can also use the Riley class, which is stored as a controlled vocabulary.

ACKNOWLEDGEMENTS

This article is dedicated to the memory of Marie-Adèle Rajandream. We would like to thank Professor Christopher Newbold (University of Oxford Institute of Molecular Medicine) for his suggestions on improving the design of the Artemis Gene Builder.

Funding: This work was supported by the Wellcome Trust through their funding of the Pathogen Genomics group at the Wellcome Trust Sanger Institute.

Conflict of Interest: none declared.

REFERENCES

Bendtsen,J.D. *et al*. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*., **340**, 783–795.

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

Carver,T.J. *et al.* (2005) ACT: the Artemis Comparison Tool. *Bioinformatics*, **21**, 3422–3423.

Eilbeck,K. *et al*. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol*., **6**, R44.

Mulder,N.J. *et al*. (2007) New developments in the InterPro database. *Nucleic Acids Res*., **35**, D224–D228.

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