

# Artemis: the Goddess of the Hunt

Nick Thomson

2002 saw the start of a new venture for the SGM and The Pathogen Sequencing Unit (PSU) of the Wellcome Trust Sanger Institute. This partnership was forged by the observation that the public DNA sequence databases have continued to expand, almost exponentially year on year. Luckily for us this trend does not look like changing and the future ahead for whole-genome sequencing appears rosy, with >350 bacterial genome projects currently in progress (see Stephen Bentley's article on p. 16). However, there is an obvious truth that comes with this volume of sequencing. Without appropriate and accessible *in silico* tools with which to meld this mass of information into a more readily interpreted form, we will only ever glance at the surface and never dig deeper into the wealth of data just waiting to be exploited within these public sequence storehouses.

To keep pace with this rate of data generation we need to embrace tools to facilitate whole-genome analysis, as well as utilize software that allows us to compare and extract information from multiple genomes. This will require a move away from trusted software such as GCG and the web-based cut and paste genre of analysis packages, to include more specialized tools.

In an attempt to address this issue we came up with a plan for a series of five one-day regional workshops which would build on our experience of whole-genome analysis. The emphasis was placed on squeezing in as much 'hands-on' experience of our genome analysis tools as was possible in one day. An enthusiastic team of demonstrators (Fig. 1) backed up the practical work, with expertise in bacterial and eukaryotic genomics (ranging from Gram-negative and -positive bugs across to malaria and on to humans and mice), as well as in relevant aspects of computing. In addition to extolling the virtues of whole-genome sequence, we also felt it important to cover some of the possible pitfalls of using these data, and so we also fitted in several short talks to deal with, for example, the differences between draft and finished sequence, automated versus manual annotation and so on.

## ● Artemis and the Artemis Comparison Tool (ACT)

The computer programs featured in these workshops, Artemis and ACT, were both developed 'in-house' by Kim Rutherford and represent the core software for the analysis of both prokaryotic and eukaryotic genomes within the PSU. Artemis is a genome viewer program, which allows the user to get away from the relatively faceless EMBL- and GenBank-style database files, or reams of printed sequence marked with a highlighter pen (based on my previous experiences), and view the genome in a graphical and highly interactive format (Fig. 2). Context is probably the most important facet of whole-genome analysis and Artemis is designed to



present multiple lines of evidence within a genomic context. This manifests itself as the ability to zoom in to look for fine DNA motifs as well as being able to zoom out and bring into view operons, several kilobases of the genome or in fact to view the entire genome in one screen. It is also possible to perform quite sophisticated analysis and store the output within the 'Artemis environment' to be accessed later. This is a real bonus for people, like myself, with a paper to desk weight ratio approaching 1:1.

Artemis has also proven to be an invaluable 'hands-on' tool for teaching concepts such as gene structure and organization at all levels:

*'Since the workshop we have incorporated Artemis and ACT into both undergraduate and postgraduate teaching. These programs provide excellent tools for active learning in terms of the investigation of microbial genomes, a field of self-evident importance though not necessarily easy to teach. The visual appeal and dynamic nature of the user interface go beyond what can be readily achieved on the Web.'*

Dr Peter Miller (University of Liverpool)

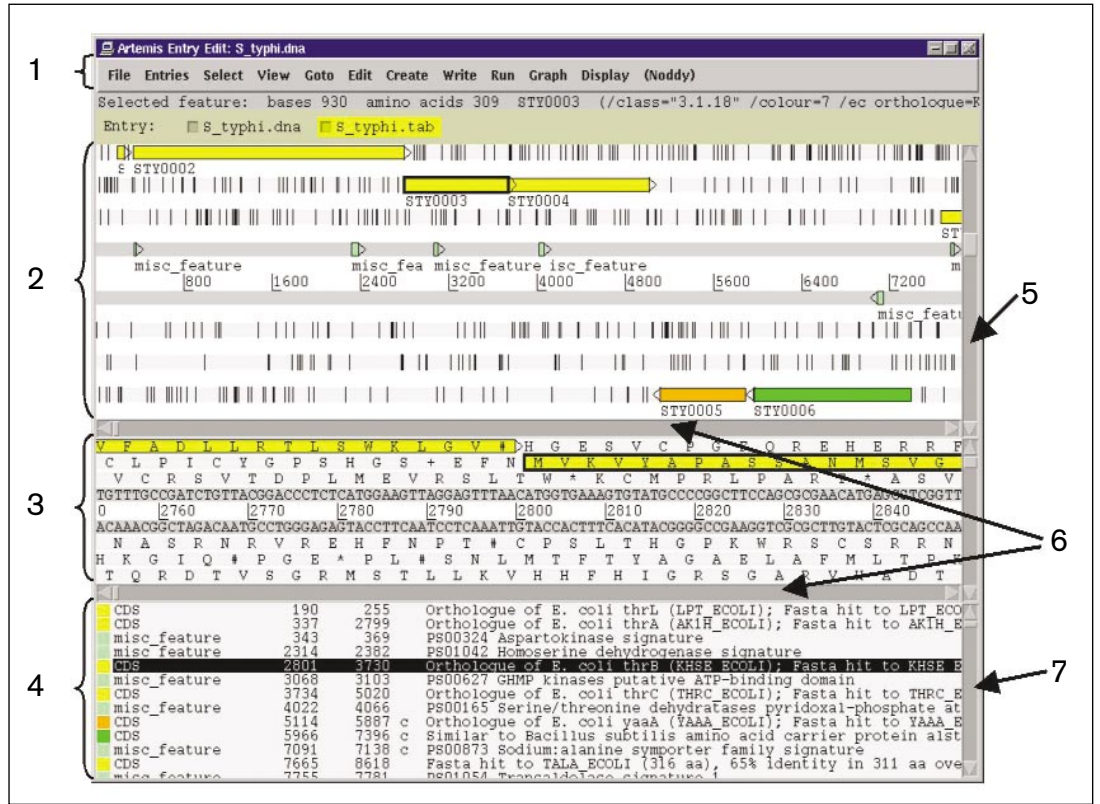
The comparative genomic tool, ACT, is essentially composed of three layers or windows (Fig. 3), depending on the number of genomes. The top and bottom layers are essentially mini Artemis windows (with their inherited functionality), showing the linear representations of the genomes with their associated features. If the upper and lower layers represent the bread, then the middle window is the meat in the sandwich, showing red blocks which span the middle layer and link conserved regions (Fig. 3). Consequently, if you were comparing two identical genome sequences you would see a solid red block extending over the length of the two sequences in the middle layer. If insertions were present in either of

Recently the SGM and the Wellcome Trust Sanger Institute have held some very successful bioinformatics workshops. Nick Thomson, one of the tutors, explains what went on.

ABOVE:  
Fig. 1. The PSU demonstrators and the hosts of the Bristol workshop. From left to right: Nick Thomson, Kim Rutherford, Mohammed Sebahia, Howard Jenkinson (SGM organizing representative), Matthew Holden, Julian Parkhill and Matthew Avison (local host). Those members of the PSU not pictured include Arnab Pain, Rhian Gwilliams, Ana Cerdêno-Tarraga.  
COURTESY N. THOMSON

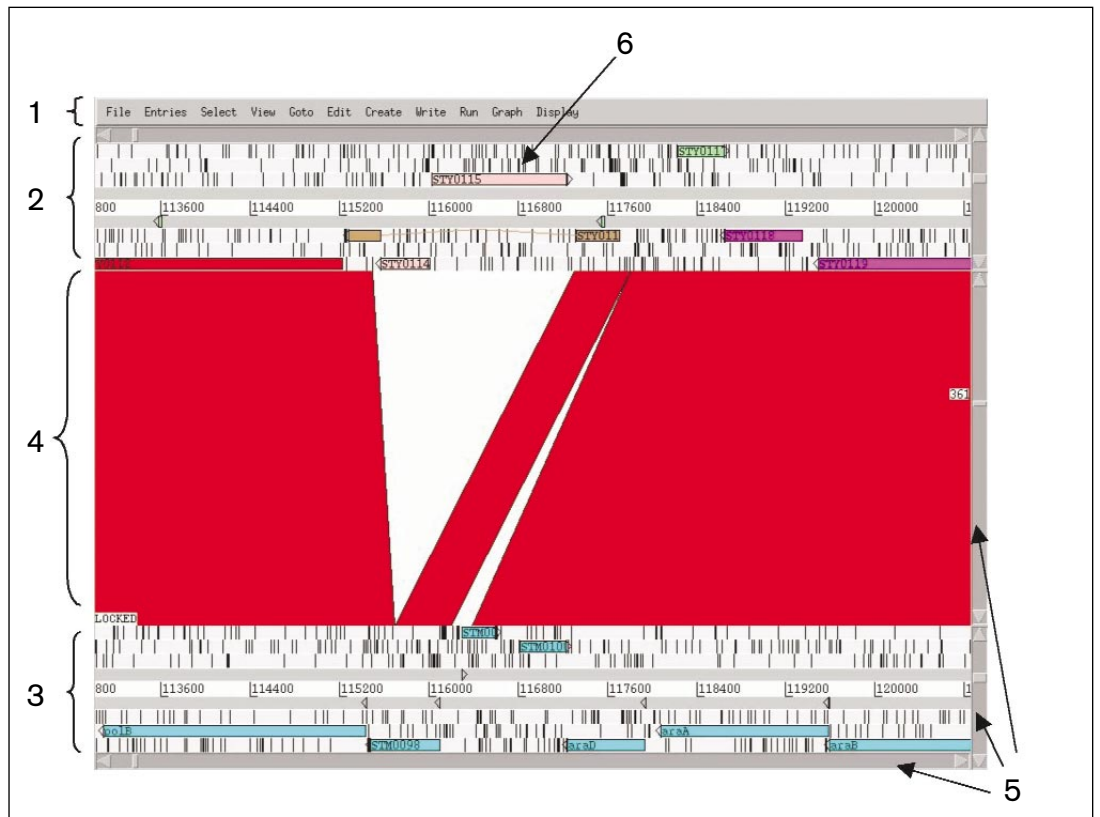
RIGHT:

**Fig. 2.** Artemis. (1) Drop-down menus. (2) Main Sequence View panel showing two central grey lines representing the forward (top) and reverse (bottom) DNA strands. Above and below the DNA strands are the three forward and three reverse reading frames. Stop codons are marked as black vertical bars. Genes and other features (e.g. Pfam and Prosite matches) are displayed as coloured boxes. (3) This is a zoomed-in view of the main panel (2) showing individual nucleotides and amino acids. (4) List of the various features in the order that they occur on the DNA as shown in the other windows. (5-7) Sliders for scrolling and zooming in/out. COURTESY N. THOMSON



RIGHT:

**Fig. 3.** ACT. (1) Drop-down menus applying to both sequence windows. (2 and 3) Artemis style view panels showing the sequences being compared. (4) Comparison layer displaying regions of conservation as red blocks. (5) Sliders that allow you to align and move along the genomes and optimize the view. (6) The insertion of an IS element in the upper sequence, disrupts a gene (brown) and appears as a break in the red comparison layer (4). COURTESY N. THOMSON





ABOVE:  
**Fig. 4.** The exercises detailed in the workshop manual were backed up with a high level of supervision from a team of demonstrators involved in the original analysis of genomes featured.  
 COURTESY N. THOMSON

the genomes, they would show up as breaks between the solid red conserved regions. Data used to draw these red blocks and link conserved regions is generated by running pairwise BLASTN, or TBLASTX, comparisons of the genomes (details of how this is done can be obtained from the ACT user manual: <http://www.sanger.ac.uk/Software/ACT/manual/>).

ACT may be simple in concept, but it has proved to be very powerful in application. This is most evident when comparing the growing number of genomes from closely related organisms. In this instance it is relatively straightforward to identify regions within each genome that have undergone insertion/deletion, frameshifts, inversions and translocation events. Regions such as these can be very telling when attempting to understand how organisms have evolved and what genomic strategy has been employed (e.g. reductionist or expansionist) for an organism to adapt to a new lifestyle.

Artemis and ACT can be freely downloaded from our web pages in several versions designed to run on different platforms (<http://www.sanger.ac.uk/Software/Artemis/> or <http://www.sanger.ac.uk/Software/ACT/>). Like the organisms themselves, Artemis and ACT are undergoing continual evolution. These changes, some of which are a direct result of feedback from SGM workshops, filter quickly into the publicly available versions. Hence there is always a developmental as well as a 'tried-and-tested' version available on our website.

### ● The running order for the workshops

The workshops themselves were split in two halves: the morning session began with an overview of whole-genome sequencing, followed by a practical session using Artemis. The afternoon followed a similar format but focused on comparative genomics. All the exercises were laid out in a comprehensive manual, written especially for these workshops, backed up by constant supervision (Fig. 4). The exercises were designed to not only demonstrate the functionality of the software, but also to highlight some of the more idiosyncratic and fascinating aspects of the bacterial genomes featured.

*'Because of its multi-layered approach: seminar, learn-as-you-try, and one-to-one problem solving, the workshop inspired and educated newcomers and experienced users, alike.'*

Dr Matthew Avison (University of Bristol)

Each of the guided exercises was written and championed by a demonstrator involved in the original analysis of that genome. Because many of the people on the courses were experts working on many of the genomes featured we were also able to pass on the baton to them and have some great discussions, undoubtedly one of the major highlights for us.

We realized that we could not make people 'power users' of Artemis and ACT in a day, but we could

hopefully generate enough enthusiasm to overcome that initial activation energy that dogs us all when getting to grips with new software, video recorders, HSE rules... etc., and we appear to have had success in this endeavour:

*'Since the workshop we have used Artemis and ACT for the annotation of the Bacillus anthracis genome and its comparison with the genomes of other members of the genus Bacillus. In particular, we have used ACT for gross topological comparisons of gene organization and to generate comparative gene organization maps in relation to genes of specific interest.'*

Professor Colin Harwood (University of Newcastle)

### ● Location, location, location

Our task to find suitable locations for these workshops was not as simple as we initially anticipated. Whilst it is easy to find Unix-based computers in most UK universities, it is very difficult to find enough of them in the same room for a large-scale computer-based workshop. In the end we found some fantastic facilities at Newcastle, Bristol and Liverpool. We also played two home fixtures back here in Cambridge. We were all made very welcome by the event hosts (Colin Harwood, Anil Wipat, Matthew Avison and Peter Miller) who put in huge amounts of time making their events run very smoothly. As for the participants, they came from all strata of academia as well as some from industry. In addition to the local attendees some travelled great distances to get to these events. At the Cambridge workshop people flew in from Aberdeen and The Netherlands.

### ● The title: 'Artemis: the Goddess of the Hunt'

Finally, to explain the title: Artemis, a character from Greek mythology, was hunter-in-chief to the gods. We felt this was an appropriate name for the genome analysis software, and consequently a banner for these workshops, because it can be used to search out the trends and minutiae hidden within a veritable forest of data. So come on, join in, the hunt is on!

● Dr Nick Thomson is a Senior Computer Biologist at the the Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, UK

If you want any more information on the SGM workshops see [www.sgm.ac.uk](http://www.sgm.ac.uk) or email us at [sgm@sanger.ac.uk](mailto:sgm@sanger.ac.uk)

More bioinformatics workshops will take place in 2003. See p. 35 for details.