MUTAGEN: Multi-User Tool for Annotating GENomes

Kim Brügger1,∗, Peter Redder1 and Marie Skovgaard2

1Danish Archaea Centre, Institute of Molecular Biology, University of Copenhagen, Salvgade 83H, DK-1307 Copenhagen K, Denmark and 2Center for Biological Sequence Analysis, The Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

Received on February 28, 2003; revised on May 30, 2003; accepted on June 30, 2003

ABSTRACT

Summary: MUTAGEN is a free prokaryotic annotation system. It offers the advantages of genome comparison, graphical sequence browsers, search facilities and open-source for user-specific adjustments. The web-interface allows several users to access the system from standard desktop computers. The Sulfolobus acidocaldarius genome, and several plasmids and viruses have so far been analysed and annotated using MUTAGEN.

Availability: MUTAGEN is released as open-source software under GPL. The code is available for download and/or contribution at http://dac.molbio.ku.dk/bioinformatics/MUTAGEN/

Contact: brugger@mermaid.molbio.ku.dk

In recent years, several research groups have published genome sequences from closely related organisms, and several systems are available for annotating these genomes. The commercial systems include MAGPIE (Gaasterland and Sensen, 1996), ERGO (Overbeek et al., 2003), and the open-source systems include Artemis (Rutherford et al., 2000), Genquire (Wilkinson et al., 2002) and GenDB (Meyer et al., 2003). However, none of these systems take full advantage of sequences available from related species.

MUTAGEN is a new annotation system that exploits features developed for earlier annotation systems, such as ORF browser, 6-frame translation browser, and data collection from several data sources. In addition MUTAGEN incorporates a genome comparison system, which facilitates both the annotation process and analyses of the whole genome. MUTAGEN was developed for the annotation of Sulfolobus acidocaldarius and a strong motivation for developing MUTAGEN was the need for data sharing within the sequencing consortium and that users would not have to install programs locally. Furthermore, tools for genome comparison were needed to utilize the information from the sequences of the closely related species Sulfolobus solfataricus (She et al., 2001) and Sulfolobus tokodaii (Kawarabayasi et al., 2001).

∗To whom correspondence should be addressed.

ANNOTATION IN MUTAGEN

In MUTAGEN, we have chosen a manual annotation method where the collection and presentation of relevant data will aid the human annotator in this otherwise time consuming task. The alternative is automatic annotation, which is fast, but suffers from loss of organism-specific knowledge and a lack of debate of questionable assignments. However, if required, MUTAGEN can be altered easily to display information derived from an automatic annotation system.

The assignment of function is not trivial for many genes and knowledge from different experts is important for the correct identification of functions (Brenner, 1999). In MUTAGEN, it is possible to store multiple annotations for a gene entry and these can all be evaluated during the final round of annotation.

In order to minimize the time spent on a sequencing project, it is important to begin the annotation before the sequence is completed. MUTAGEN links annotations from one sequence assembly to a more advanced one by aligning all genes in the assemblies. Information from the previous assemblies is thereby retained. This simple approach solves a problem experienced with other annotation systems, where the user had to re-find genes because they had been renamed during sequence updates.

The database is organized such that it permits annotation and comparison of several genomes. This gives the rare possibility to compare both individual genes and their organization in operons in related species. All genes from all the genomes in the database are extracted and aligned to each other with BLAST (Altschul et al., 1997). Gene pairs with an E-value of less than 10e-6 and a full-length match of the query sequence are defined as homologues and this information is stored in a database. It is possible to view the homologous genes as multiple alignments generated with T-Coffee (Notredame et al., 2000) and/or as graphical representations of the genomic neighbourhood where all homologous genes are shown with colours (Fig. 1).

Username and password are required to gain access to a session on the MUTAGEN server, thus protecting unpublished
MUTAGEN: Multi-user tool for annotating GENomes

Fig. 1. The genomic neighbourhood (±10 kbp) of an ORF in *S.tokodaii* and *S.solfataricus* (opposite strand orientation). The homology browser shows each set of homologous genes with the same colour and pattern (here in greyscale), and reveals a conserved gene order in the region. White rectangles represent ORFs without homologues.

Sequence data on the World Wide Web. User-input is sanity-checked to prevent misuse. Additionally, the different functionalities in the system belong to separate groups. To avoid local program installations by users, both annotation, and most maintenance, is done via a web-interface.

IMPLEMENTATION

MUTAGEN has been developed for the annotation of the archaeal genome *S.acidocaldarius* [http://dac.molbio.ku.dk/](http://dac.molbio.ku.dk/). It has also been used for several plasmids and viruses.

MUTAGEN is a collection of PERL scripts linked to specific libraries. Each library is responsible for a unique functionality. This organization is convenient for adjustments of the system and for addition of new features and functionalities. Data are stored in a MySQL database. The data collection modules can be modified easily and expanded to include local resources.

The system has, so far, been tested on Linux and Solaris workstations. The web interface has been used on many browsers including Netscape and Internet Explorer on Windows, Linux and Macintosh platforms.

ACKNOWLEDGEMENTS

We wish to thank L. Chen for providing the *S.acidocaldarius* genome sequence, and to R.A. Garrett and A. Krogh for support and discussions. K.B. is funded by a grant from the Danish Science Research Council for Centre for Genomics, Biodiversity and Evolution of Archaea. P.R. is funded by a PhD stipend from the University of Copenhagen. M.S. is funded by EU Cell Factory Project, SCREEN, QLK3-CT-2000-00649.

REFERENCES


Kawarabayasi,Y., Hino,Y., Horikawa,H., Jin-no,K., Takahashi,M., Sekine,M., Baba,S., Ankai,A., Kosugi,H., Hosoyama,A. *et al.* (2001) Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain 7. *DNA Res.*, 8, 123–140.


