Published online 6 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D413–D415***
doi:10.1093/nar/gkl847

The *Sulfolobus* database

Kim Brügger*

Institute of Molecular Biology, University of Copenhagen, Sølvgade 83H, DK1307 Copenhagen K, Denmark

Received August 10, 2006; Revised October 8, 2006; Accepted October 9, 2006

ABSTRACT

The *Sulfolobus* database (http://www.sulfolobus.org) integrates, for the first time, all currently available *Sulfolobus* chromosome sequences with annotations. It also includes all the sequence data for the extrachromosomal elements which can propagate in *Sulfolobus* organisms. All genomes and annotations deposited in GenBank are included in the database and a genefinder has been run on the sequences to ensure that all potential genes are present, and identifiable, in the database. Every month, all genes are searched against a range of external databases and new results are incorporated. The *Sulfolobus* database was developed as an asset to the rapidly-growing international community working with *Sulfolobus* as a model organism for the kingdom Crenarchaeota of the Archaea. It was accessed more that 46 000 times in its first year. The database aims to provide researchers easy access to sequence and gene information and the web-interface includes various searches, free text and BLAST, as well as genome browsing and data extraction. Updated annotations are incorporated regularly and the database will continue to expand as new information becomes available. This includes new sequences, newly identified genes, annotations and other related information.

INTRODUCTION

Organisms from the *Sulfolobus* genus have been selected by the international research community as model organisms for investigating the biology of the Crenarchaeota, which is one of the two major kingdoms of the Archaea. Organisms belonging to the *Sulfolobus* genus are all aerobic and grow optimally around 80°C and pH 2–4. Much of our current knowledge of archaeal and crenarchaeal mechanisms involved in the cell cycle (1), DNA replication (2), DNA repair (3) and RNA processing (4), derive from studies on *Sulfolobus* species.

The database currently contains three fully sequenced *Sulfolobus* chromosomes [*Sulfolobus acidocaldarius* (5), *Sulfolobus solfataricus* (6) and *Sulfolobus tokodaii* (7)] as well as many sequenced extrachromosomal elements, plasmids and viruses, which can propagate in *Sulfolobus*. The plasmids are either cryptic or conjugative, and the viruses have been classified into seven new viral families (8). By including the extrachromosomal elements, genome comparison can be made between the extrachromosomal elements and chromosomes. Sequence comparison between extrachromosomal elements is a very useful approach for understanding how they evolve, and which genes are mandatory for propagation and spreading of these elements. Furthermore, it is useful for identifying and analysing the integration of extrachromosomal elements into genomes.

New *Sulfolobus* sequences provided by the international community will be incorporated into the database as they become available. We are currently finishing two genomes; *Sulfolobus islandicus* and *Acidianus brierleyi*. The latter is phylogenetically close to *Sulfolobus* although it is anaerobic. Eight additional *S.islandicus* genomes are being sequenced (9), and these will be included when available. Moreover, several newly isolated viruses are being sequenced in our laboratory, all of which will be integrated into the database.

New genes, corrections and/or new annotations will be added as this information becomes available, or upon request from other researchers. By providing these services we will ensure that all the latest corrections to annotations and functional assignments of proteins previously classified as hypothetical are available in one place, such that researches do not have to search several databases for this information.

MATERIALS AND METHODS

MUTAGEN (10) was originally developed as an annotation tool for the *S.acidocaldarius* genome, but after the genome was published, the system was further developed into a sequence database and made publicly available. The database is backed by a relational database which can be accessed through a simple, yet comprehensive, web interface. This gives the users the opportunity to perform mining and visualization of the information in the database. The idea with the *Sulfolobus* database is that it should be easy to use, yet powerful enough to perform advanced analyses of the data it contains.

All aspects of the system are free to the public. Moreover, all the data in the database can be extracted, either fully

*To whom correspondence should be addressed. Tel: +45 35 32 20 18; Fax: +45 35 32 20 40; Email: brugger@mermaid.molbio.ku.dk

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

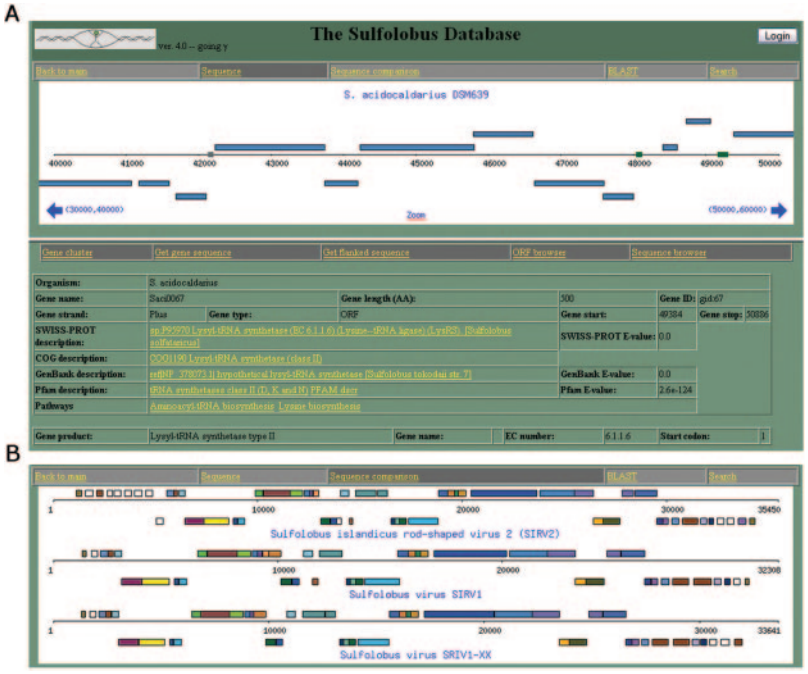

Figure 1. The database has a set of basic pages that makes it possible to navigate and display database entries in various ways. (**A**) This is the main sequence browser where the user can select genes and display information based on searches against external databases. (**B**) A full alignment of 3 different viruses. Homologous genes are coloured identically to easily identify conserved genes and operons. A similar picture can be obtained showing the genetic neighbourhood of genes between all sequences in the database.

or partially, in various formats. The whole system with programs and database can be acquired upon request, providing the possibility for other laboratories to install the system locally with an additional selection of their own sequences.

The main sequence browser shows all genes and features which are annotated in the selected sequence (Figure 1a). Features include repeated regions, and other structurally interesting genomic regions. First the user selects which sequence should be displayed. It is always possible to zoom in and out, or even to select a different region of the sequence to be displayed. The user can then select either a gene, or a feature, by clicking on them. Data associated with the selected gene will be displayed, including the name and position of the gene, and which strand it is encoded on. The annotation originating from the GenBank file is displayed together with results from pre-computed analyses and searches against various databases. The user has the possibility to extract the gene, either as DNA or protein sequence, or extract the gene with the surrounding flanking sequence. Moreover, it is possible to open a new browser which displays the sequence at a DNA level for analysis of upstream and downstream regions or to check that a correct start codon has been assigned. Currently the sequences are analysed against GenBank (11), COG (12), SwissProt (13), pfam (14) and KEGG pathways (15). For the latter, a pathway map can be displayed where all the genes involved in the pathway are highlighted and can be accessed in the database. Potential transmembrane helices and signal peptides are predicted by utilising TMHMM (16) and signalp (17), respectively. When possible the results give access to detailed information by displaying relevant pages provided by the externally public databases (e.g. Swissprot, Genbank, &c) the data originated from. Since comparative analyses of both genes and genomes are dependent on the data being regularly updated, all the pre-computed searches against external databases are updated monthly. This ensures that the newest available data are accessible for the users.

Families of homologous proteins provide an important basis for evaluating conservation and, gain or loss of gene function in a genome, especially when these data can be utilized to visualize the genomic neighbourhood of genes of interest. These data are made accessible by identifying all homologous genes in the database using tribeMCL (18). Thus, orthologous and paralogous protein genes can readily be identified both within a given sequence and also with in the rest of the database. They can also be supplemented by showing a graphical alignment of genes surrounding a gene of interest, or if working with the extrachromosomal elements, by graphically aligning the full sequence (Figure 1b). In this way, it is possible to identify and analyse genomic neighbourhoods showing conserved operons and sequence regions. Moreover, this is especially useful when analysing the evolution of plasmids and viruses or the occurrence of chromosomally integrated elements. Thus, it is also possible to make multiple alignments of all, or a selection of the homologous protein genes, or to export the protein sequences for local computations of ones interest.

The web interface provides a set of forms through which the user can easily query the annotation and pre-computed analysis data. Genes of interest can be identified by performing text searches in the imported annotations and/or data obtained from the external database searches, making it possible to identify easily all genes with a specific domain or annotated function. It is also possible to perform BLAST

searches against sequences in the *Sulfolobus* database, searching both DNA sequence and annotated proteins. All new published prokaryotic microbial genomes are added to the database automatically, so the user can compare the *Sulfolobus* sequences with other micro organisms. When querying with protein sequences, the user has the possibility to access the sequence browser along with relevant information, as described above.

DISCUSSION

The *Sulfolobus* database is a powerful research tool that has been the cornerstone in our genome analyses. It has enabled us to compare and analyse sequences from newly sequenced genomes, independently of whether the sequences are chromosomal, plasmid or viral. The ability to readily identify homologous genes and to compare the gene orders facilitates the identification of potentially important genes and operons. Furthermore, it is straightforward to make genomic comparisons between any, or all, of the genomes in the database, revealing conserved gene clusters, genetic neighbourhoods etc.

The operating principles of the *Sulfolobus* database are to assist and respond to the users' needs, and to capitalize on the major and vast efforts of the *Sulfolobus* research community, by providing an important and useful data resource. Most important, the database is interactive and has been established as a central facility where researchers can add, correct and/or update gene annotations, thereby ensuring that it will be the site with the most up-to-date annotations and gene predictions for the rapidly growing *Sulfolobus* community.

ACKNOWLEDGEMENTS

The database runs on hardware supported by grants from the Danish Research Council for Natural Science. The author was supported by grants from Copenhagen University and the Danish Science Research Council. Funding to pay the Open Access publication charges for this article was provided by the University of Copenhagen.

Conflict of interest statement. None declared.

REFERENCES

1. Lundgren,M. and Bernander,R. (2005) Archaeal cell cycle progress. *Curr. Opin. Microbiol.*, **8**, 662–668.
2. Duggin,I.G. and Bell,S.D. (2006) The chromosome replication machinery of the archaeon *Sulfolobus solfataricus*. *J. Biol. Chem.*, **281**, 15029–15032.
3. White,M.F. (2003) Archaeal DNA repair: paradigms and puzzles. *Biochem. Soc. Trans.*, **31**, 690–693.
4. Grote,M., Dijk,J. and Reinhardt,R. (1986) Ribosomal and DNA binding proteins of the thermoacidophilic archaebacterium *Sulfolobus acidocaldarius*. *Biochim. Biophys. Acta.*, **873**, 405–413.
5. Chen,L., Brugger,K., Skovgaard,M., Redder,P., She,Q., Torarinsson,E., Greve,B., Awayez,M., Zibat,A., Klenk,H.P. *et al.* (2005) The genome of *Sulfolobus acidocaldarius*, a model organism of the Crenarchaeota. *J. Bacteriol.*, **187**, 4992–4999.
6. She,Q., Singh,R.K., Confalonieri,F., Zivanovic,Y., Allard,G., Awayez,M.J., Chan-Weiher,C.C., Clausen,I.G., Curtis,B.A., De Moors,A. *et al.* (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc. Natl Acad. Sci. USA*, **98**, 7835–7840.
7. 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* strain7. *DNA Res.*, **8**, 123–140.
8. Prangishvili,D. and Garrett,R.A. (2005) Viruses of hyperthermophilic Crenarchaea. *Trends Microbiol.*, **13**, 535–542.
9. Liolios,K., Tavernarakis,N., Hugenholtz,P. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.*, **34**, D332–D334.
10. Brügger,K., Redder,P. and Skovgaard,M. (2003) MUTAGEN: multi-user tool for annotating genomes. *Bioinformatics*, **19**, 2480–2481.
11. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. *Nucleic Acids Res.*, **34**, D16–D20.
12. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
13. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
14. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
15. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
16. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
17. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
18. Enright,A.J., Kunin,V. and Ouzounis,C.A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.*, **31**, 4632–4638.