



HAL
open science

BAGET 2.0: an updated web tool for the effortless retrieval of prokaryotic gene context and sequence

Benjamin Hepp, Violette da Cunha, Florence Lorieux, J. Oberto

► **To cite this version:**

Benjamin Hepp, Violette da Cunha, Florence Lorieux, J. Oberto. BAGET 2.0: an updated web tool for the effortless retrieval of prokaryotic gene context and sequence. *Bioinformatics*, 2021, 37 (17), pp.2750-2752. 10.1093/bioinformatics/btab082 . hal-03366532

HAL Id: hal-03366532

<https://hal.science/hal-03366532v1>

Submitted on 5 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Genome analysis

BAGET 2.0: an updated web tool for the effortless retrieval of prokaryotic gene context and sequence

Benjamin Hepp, Violette Da Cunha, Florence Lorieux and Jacques Oberto  *

Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on December 2, 2020; revised on January 12, 2021; editorial decision on January 27, 2021; accepted on January 29, 2021

Abstract

Motivation: The retrieval of a single gene sequence and context from completely sequenced bacterial and archaeal genomes constitutes an intimidating task for the wet bench biologist. Existing web-based genome browsers are either too complex for routine use or only provide a subset of the available prokaryotic genomes.

Results: We have developed BAGET 2.0 (Bacterial and Archaeal Gene Exploration Tool), an updated web service granting access in just three mouse clicks to the sequence and synteny of any gene from completely sequenced bacteria and archaea. User-provided annotated genomes can be processed as well. BAGET 2.0 relies on a local database updated on a daily basis.

Availability and implementation: BAGET 2.0 befits all current browsers such as Chrome, Firefox, Edge, Opera and Safari. Internet Explorer 11 is supported. BAGET 2.0 is freely accessible at <https://archaea.i2bc.paris-saclay.fr/baget/>

Contact: jacques.oberto@i2bc.paris-saclay.fr

1 Introduction

The completely sequenced prokaryotic genomes from the National Center for Biotechnology Information (NCBI) constitute an indispensable and freely available source of information for a number of biological disciplines. Several desktop programs provide local access to nucleotide-level information from complete genomes but require installation in addition to the download, formatting and storage of large genomic files. These include the free BAC-BROWSER (Garanina *et al.*, 2018) and the commercially available SnapGene (www.snapgene.com) or Geneious (Kearse *et al.*, 2012). To overcome these limitations, several web-based genome browsers such as Gbrowse (Stein, 2013), Ensembl (Yates *et al.*, 2020) and UCSC Genome Browser (Lee *et al.*, 2020) have been developed but are devoted exclusively to eukaryotic genomes. The UCSC Microbial Genome Browser (Chan *et al.*, 2012), EnsemblGenomes (Howe *et al.*, 2020) and MicrobesOnline (Dehal *et al.*, 2010) provide only a limited number of prokaryotic genomes or exclusively bacterial ones and are not updated on a regular basis. The NCBI microbial genome browser (www.ncbi.nlm.nih.gov/genome/microbes/) provides to most comprehensive access to prokaryotic genomes. However, with a searchable organism list over 5000-page-long and a complex horizontal genome viewer, this resource requires a number of user gestures and its results are difficult to export. The retrieval of a particular gene and its immediate synteny from a given prokaryotic genome remains a daunting task for wet bench experimentalists. For this reason, we developed the BAGET web service to provide a free, immediate and effortless access to the sequence, genomic context, intergenic regions and coding capacity of any gene from all complete

prokaryotic genomes (Oberto, 2008). Here, we propose the BAGET 2.0 update which exhibits additional features, significant performance improvements and the capacity to handle and provide effortless access to >25 000 prokaryotic genomes.

2 Materials and methods

BAGET 2.0 has been completely rewritten in C# 8.0 and .NET Core 3.1 to comply with the industry-standard Model-View-Controller architectural framework and the latest HTML 5.0 specifications. The use of JQuery (jquery.com) and D3 (d3js.org) JavaScript libraries enables asynchronous data transfer and access to the HTML 5.0 graphical canvas, respectively. The BAGET 2.0 database is hosted locally and shared with the FITBAR and SYNTTAX web services (Oberto, 2010, 2013). Incremental database updates occur daily by retrieving new GBFF genomes from the NCBI ([ftp.ncbi.nlm.nih.gov](ftp://ftp.ncbi.nlm.nih.gov)) (Sayers *et al.*, 2020). Multiple replicons present in some genomes are accessible separately in BAGET 2.0 and identified by the organism name tag followed by the C1, C2 ... Cn suffixes, by decreasing size.

3 Features

All the features provided by the original BAGET web service have been maintained and several additions and improvements were implemented in BAGET 2.0 as follows:

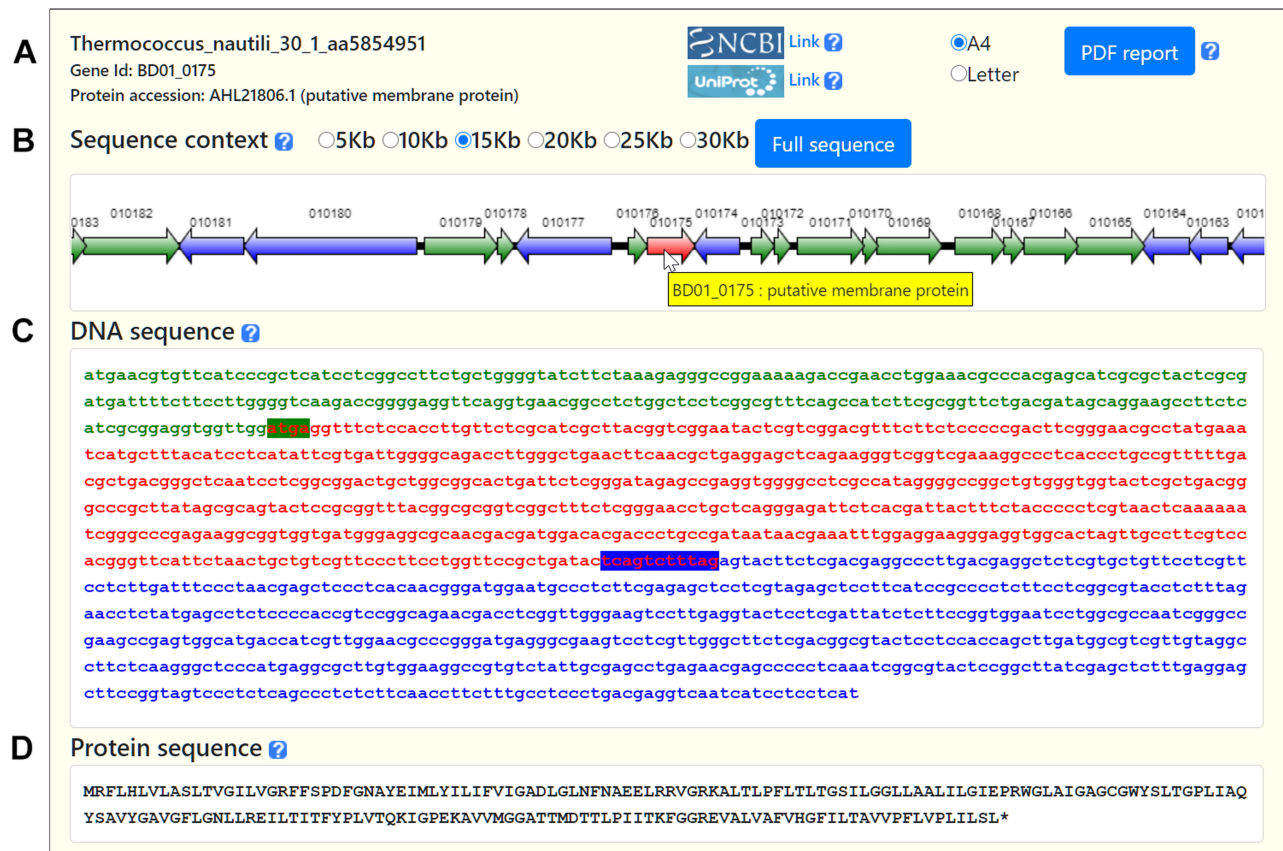


Fig. 1. BAGET 2.0 report for the BD01_0175 gene from *T.nautili* 30-1. (A) Information relative to the selected gene. (B) Graphical context map spanning a user-selectable interval. The gene of interest is shown in the center of the map. (C) The DNA sequence of the BD01_0175 open reading frame is shown in red. The extent of gene overlapping with the preceding and following genes is identified by text background color. (D) AHL21806.1 protein sequence in amino acid single letter code

- To facilitate user interaction, the initial replicon selection from the database is based on the first letter of the corresponding organism name. A selectable scrolling list of organisms with the same initial is then provided.
- BAGET 2.0 accepts the submission of user-provided annotated genomes in the GenBank GBFF format. Uploaded genomes are processed in real time and their features are exposed similarly to database entries. User-uploaded genomes are not added to the database and remain confidential.
- Upon selection or user-uploading of a replicon, a corresponding gene list is created where individual genes can be selected to generate a complete report.
- Text queries on the annotation of uploaded or database genomes provide a user-selectable gene list.
- Gene selection generates a proportionally scaled color-coded context map covering a selectable genomic interval between 5 and 30 kb. The gene of interest is depicted in red at the center of the map in 5' to 3' orientation. Surrounding genes in the same orientation are shown in green while the antiparallel genes, in blue. Information on mapped genes is accessible by mouse hovering. Navigation through mapped genes is accessible by mouse clicking.
- The DNA sequence of the gene of interest and the immediately surrounding genes is shown using the same color-code and genetic orientation as the context map. For protein-coding genes, the colored sequences refer to open reading frames. The intergenic regions are shown in black and gene overlaps are displayed

using text background colors. The protein sequence of coding genes is displayed in single letter amino acid code.

- Each selected coding gene is provided with an external link to the respective protein entry at the NCBI (Sayers et al., 2020) and at Uniprot (UniProt, 2021).
- PDF reports can be generated either in US-letter or DIN A4 format and downloaded.
- A contextual help file is accessible through blue question mark icons.

4 Results

BAGET 2.0 provides a fast and effortless access to every gene from over 54 000 replicons corresponding to >25 000 bacterial and archaeal genomes. Its user interface consisting of a single page loaded only once at startup does not require navigation through numbers of forms. In order to mimic user-friendly desktop applications, BAGET 2.0 exchanges data with the server in asynchronous mode with a significant increase in performance. Only three mouse clicks were required to access the BD01_0175 gene sequence and genomic context from *Thermococcus nautili* 30-1 (Fig. 1). The major assets of this web service reside in its user friendliness, daily updated database and quality of the exportable results. To ensure user confidentiality, submitted data, generated results and PDF reports are never saved to disk but reside exclusively in volatile memory for the duration of the user session. BAGET 2.0 does not track users nor use cookies. The versatility and user-friendliness of BAGET 2.0 will benefit wet bench experimentalists and phylogeneticists alike.

Funding

This work was funded by the 'Centre Nationale de la Recherche Scientifique, Agence Nationale de la Recherche (grant ANR-19-CE11-0007).

Conflict of Interest: none declared.

References

- Chan,P.P. *et al.* (2012) The UCSC archaeal genome browser: 2012 update. *Nucleic Acids Res.*, **40**, D646–D652.
- Dehal,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.
- Garanina,I.A. *et al.* (2018) BAC-BROWSER: the tool for visualization and analysis of prokaryotic genomes. *Front. Microbiol.*, **9**, 2827.
- Howe,K.L. *et al.* (2020) Ensembl genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.*, **48**, D689–D695.
- Kearse,M. *et al.* (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.
- Lee,C.M. *et al.* (2020) UCSC genome browser enters 20th year. *Nucleic Acids Res.*, **48**, D756–D761.
- Oberto,J. (2008) BAGET: a web server for the effortless retrieval of prokaryotic gene context and sequence. *Bioinformatics*, **24**, 424–425.
- Oberto,J. (2010) FITBAR: a web tool for the robust prediction of prokaryotic regulons. *BMC Bioinform.*, **11**, 554.
- Oberto,J. (2013) SyntTax: a web server linking synteny to prokaryotic taxonomy. *BMC Bioinform.*, **14**, 4.
- Sayers,E.W. *et al.* (2020) GenBank. *Nucleic Acids Res.*, **48**, D84–D86.
- Stein,L.D. (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief. Bioinform.*, **14**, 162–171.
- UniProt,C. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
- Yates,A.D. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.