BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins

Auke J. van Heel†, Anne de Jong†, Chunxu Song, Jakob H. Viel, Jan Kok and Oscar P. Kuipers*

Molecular Genetics, GBB, University of Groningen, Groningen, 9747AG, the Netherlands

ABSTRACT

Interest in secondary metabolites such as RiPPs (ribosomally synthesized and posttranslationally modified peptides) is increasing worldwide. To facilitate the research in this field we have updated our mining web server. BAGEL4 is faster than its predecessor and is now fully independent from ORF-calling. Gene clusters of interest are discovered using the core-peptide database and/or through HMM motifs that are present in associated context genes. The databases used for mining have been updated and extended with literature references and links to UniProt and NCBI. Additionally, we have included automated promoter and terminator prediction and the option to upload RNA expression data, which can be displayed along with the identified clusters. Further improvements include the annotation of the context genes, which is now based on a fast blast against the prokaryote part of the UniRef90 database, and the improved web-BLAST feature that dynamically loads structural data such as internal cross-linking from UniProt. Overall BAGEL4 provides the user with more information through a user-friendly web-interface which simplifies data evaluation. BAGEL4 is freely accessible at http://bagel4.molgenrug.nl.

INTRODUCTION

BAGEL4 is a web server that enables users to identify and visualize gene clusters in prokaryotic DNA involved in the biosynthesis of Ribosomally synthesized and Post translationally modified Peptides (RiPPs) and (unmodified) bacteriocins. Interest in these classes of molecules is increasing due to the need for novel antibiotics and to their important role in food preservation, microbial ecology and plant biocontrol. The post translational modifications in RiPPs expand the natural structural diversity beyond the 20 genetically encoded amino acids (1). These modifications often stabilize the peptides, making them more resistant to heat and proteases.

Mining web servers help researchers to analyse the genetic potential of strains. They can aid in pinpointing the genetic origin of an observed antimicrobial activity and hence identifying the associated chemical structure (2,3). Alternatively the data can provide a starting point for (heterologous) production of novel ribosomally synthesized antimicrobial compounds (4). Also these servers have added value in annotation pipelines (5–7). Since the development of the first version of BAGEL (8) more web servers have been developed such as antiSMASH (9), PRISM (10) and recently RIPPMiner (11). They all depend on information from literature and databases such as bactibase (12) CAMPR3 (13) and the MiBig data repository (14).

Here, we present the latest version of the BAGEL suite, BAGEL4. New features such as the integration of RNA-Seq data, an improved web blast and integration of promoter and terminator predictions have been added. The databases have been thoroughly updated and we commit to supporting and maintaining the web server for years to come.

MATERIALS AND METHODS

Description of the software

BAGEL4 operates according to the flowchart depicted in Figure 1. The required input is fasta formatted DNA. Multiple files and multiple records per file are allowed; a maximum of 50 Mb can be uploaded. Alternatively a genome can be selected from a list containing full and WGS genomes (using the name or the RefSeq Accession number). DNA will only be analysed if its read length is above the set minimum (default 3000 bp). Optionally, RNA expression data can be uploaded in BedGraph track format (.BED and .BEDGRAPH extensions are allowed). The BedGraph file should contain both strands in one file. Larger datasets can be handled in consultation on our server or, alternatively, a stand-alone version is available.

The DNA is translated into six large proteins (one for each reading frame). To limit the amount of data that has

†To whom correspondence should be addressed. Tel: +31503632093; Email: o.p.kuipers@rug.nl

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2018. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
to be screened, translation is only started after a legal start
codon (ATG, GTG and TTG). Subsequently, the proteins
are screened for the (co) occurrence of certain protein mo-
tifs (Supplementary Table S1) and blasted against the core
peptide database. Based on these results so called Area(s)
Of Interest (AOI) are selected. Next, overlapping AOIs are
combined.

Once an AOI has been determined it is analyzed in de-
tail. The ORFs in the AOI are first called using Glimmer3
(15). The pipeline is setup in such a way that Glimmer3
makes a model for every defined AOI. Subsequently, small
ORFs (sORFs) are called in the intergenic regions. The de-
fault setting for these sORFs is a minimum length of 72
bp (24 amino acid residues); an overlap of 10 bp with Glim-
mer3 ORFs is allowed. All (small) ORFs are then blasted
(16) against the annotation database and the core peptide
database. If homology is found in the core peptide database,
an alignment is produced. Then promoters and terminators
are predicted (see new features for more detail).

Finally, an overview of the results is generated with links
to detailed reports per AOI. The detailed report consists of
a graphical visualisation (Figure 2) of the (annotated) genes,
promoters and terminators. Gene expression data will also
be visualised if an optional RNA-Seq file had been up-
loaded (Figure 2). Additionally, an alignment is shown if
homology with a record in the core peptide database was
found (Figure 3). UniProt structural data (cross linking,
modified residues) of this database record will also be dis-
played, if available (Figure 3).

**NEW IN BAGEL4**

**Improved core peptide (web) BLAST**

The webblast feature executes a BLAST (16) search against
the selected database and, whenever hits are found, it
visualizes an alignment that includes post-translational-
modifications. This information is imported weekly from
the public UniProt database (17). In this way, basic struc-
tural information based on homology is provided, which
gives insight into structural relatedness (cross-linking pat-
terns). Background information (literature) is provided on
the basis of the database hit. Next to being incorporated
in the full pipeline, this feature is also offered as a separate

**Six frame translation of input DNA for improved speed**

BAGEL4 translates all DNA into six large proteins, one for
each reading frame and using only legal start codons. This
converted input is used to look for motifs and core peptides.
This strategy eliminates the need for a Glimmer run at this
stage, saving computational time. Being efficient in selecting
the AOIs is important when screening large datasets.

**Integration of RNA-Seq data**

The BAGEL4 input can optionally be extended with RNA
sequencing data. This data must be supplied in the so called
BedGraph format with both strands in one file. The data
is visualized in reads per kilobase million (RPKM) below
the identified gene cluster (Figure 2). The RNA sequencing
Figure 2. Example graphics produced by BAGEL4. On the left top items can be turned on or off by clicking on the item. Genes are indicated as arrows and additional information is displayed (including a link to BLAST the protein) by mouse-over. The additional information disappears by clicking on the gene. Bottom (in blue), RNA expression data displayed in RPMK.

Figure 3. Example output of an alignment based on a BLAST hit with the core peptide database. The query is linked to a record (Nukacin A) in the core peptide database. Based on the UniProt identifier of the hit in the database, information available on modifications and bridging patterns is displayed. The leader peptide of the database record is highlighted in dark gray and modified residues are indicated with asterisks. Users should be aware that this information is only indicative for the query sequence.

data is not used for gene cluster identification but it offers two main advantages. First, it can help identifying the core peptide, if that has not been predicted by the software. Secondly, it allows examining whether the gene cluster is expressed under the condition tested.

**Integrated promoter and terminator prediction**

BAGEL4 now predicts promoters and terminators within the AOI(s). Promoter prediction is based on a DNA binding motif recognized by RNA-polymerase $\sigma^A$ (TTGACAN$^{16-18}$TATAAT). The position frequency matrix (PFM) of this motif was built using known promoter binding sites of *Escherichia coli* and *Bacillus subtilis*. This tool is also available separately in the Genome2D web server (http://genome2d.molgenrug.nl/index.php/prokaryote-promoters).

TranTermHP terminator prediction (18) was additionally implemented in the BAGEL4 pipeline. The predictions are visualised in the gene cluster (Figure 2) and can help understanding gene regulation and manually identifying core peptides.
Improved and adaptable graphical report

For each AOI a gene cluster graphic is generated that can be modified by the user. Promoters, terminators, gene names and small ORFs can be turned on or off. Moving the mouse over a certain gene provides background information and a link to BLAST its encoded protein.

DISCUSSION

The goal of BAGEL4 is to provide its users with as much information as possible on identified AOIs and to improve the annotation of novel bacterial genome sequences. Finding associated core peptides can be especially challenging when there is no homology to described compounds. Coupling different information sources can be very helpful in this respect. The new web blast feature provides users with a quick insight in the potential impact sequence differences can have on internal bridging patterns. It also often gives insight in the position of the leader cleavage site. The coupling to RNA-Seq data allows going beyond defining the genetic potential of a strain. With the increasing availability of RNA-Seq data this provides a useful additional feature. Discovery of AOIs by BAGEL4 is now fully independent of ORF calling, which has two main advantages. It firstly improves the speed of the evaluation. Secondly, not depending on ORF calling lowers the risk of missing an AOI. Overall BAGEL4 has been updated and extended with new features; it is user-friendlier and offers reliable, fast and convenient mining of bacteriocins and RiPPs.

DATA AVAILABILITY

BAGEL4 is freely available at http://bagel4.molgenrug.nl for files up to 50 mb.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We especially thank users providing feedback, now and in the future. They are invaluable in the development and improvement of software. We thank Martijn Herber for his technical knowledge and support. Special thanks to Zhibo Li and Lu Zhou for their valuable suggestions.

FUNDING

EU Horizon 2020 Rafts4Biotech (grant agreement No 720776) project (to A.J.vH.); Dutch NWO-TTW program Back to the Roots (to C.S.); Netherlands Organisation for Scientific Research [NWO, ALWOP.214 to J.H.V]. Funding for open access charge: University of Groningen.

Conflict of interest statement. None declared.

REFERENCES