W278–W281 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* *Published online 21 May 2018*
doi: 10.1093/nar/gky383

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

Received February 16, 2018; Revised April 19, 2018; Editorial Decision April 28, 2018; Accepted May 17, 2018

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 **Ri**bosomally synthesized and **P**ost translationally modified **P**eptide**s** (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.

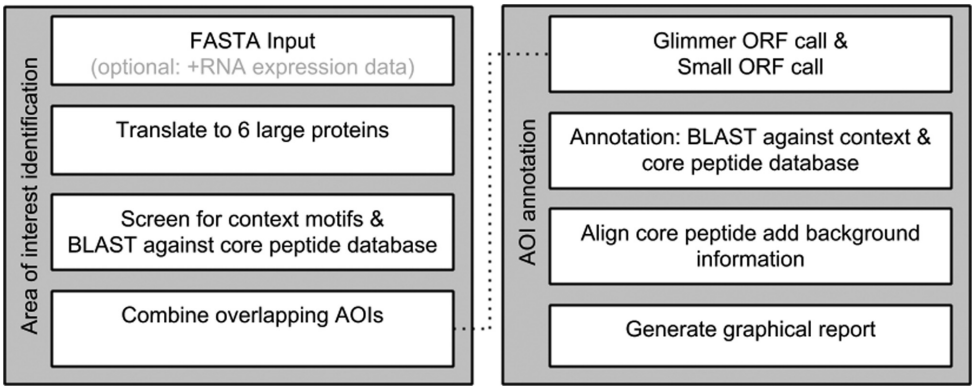

Figure 1. General flow of the BAGEL4 web server. The left part is executed for all input DNA larger than a set threshold (default 3000 bp), the right part is executed for every detected area of interest (AOI).

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 motifs (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 detail. 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 default setting for these sORFs is a minimum length of 72 bp (24 amino acid residues); an overlap of 10 bp with Glimmer3 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 uploaded (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 displayed, if available (Figure 3).

BAGEL4 databases

Core peptide database. The bacteriocin database has been updated and now contains almost 500 RiPPs (class 1, see Supplementary Figure S1), 230 unmodified bacteriocins (class 2) and 90 large (>10 kD) bacteriocins (class 3). Most records contain a link to NCBI or UniProt. Next to literature in general specific resources have been used to update our records such as RippMiner (11) and the MIBiG data repository (14). The database is available on http://bagel4.molgenrug.nl.

Annotation database. The database includes the prokaryotic part of the uniref90 database extended with the context protein database used by BAGEL3 (19). It was scanned for protein domains that are common in the context of RiPPs and this information was added to the database records.

Validation of BAGEL4

The software was validated as described previously (19). In short, the successful detection of known gene clusters was verified, while 50 recently published genomes were analysed to check for new compounds and for the appearance of false positives.

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 structural information based on homology is provided, which gives insight into structural relatedness (cross-linking patterns). 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 web-blast option (http://bagel4.molgenrug.nl/blast.php).

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

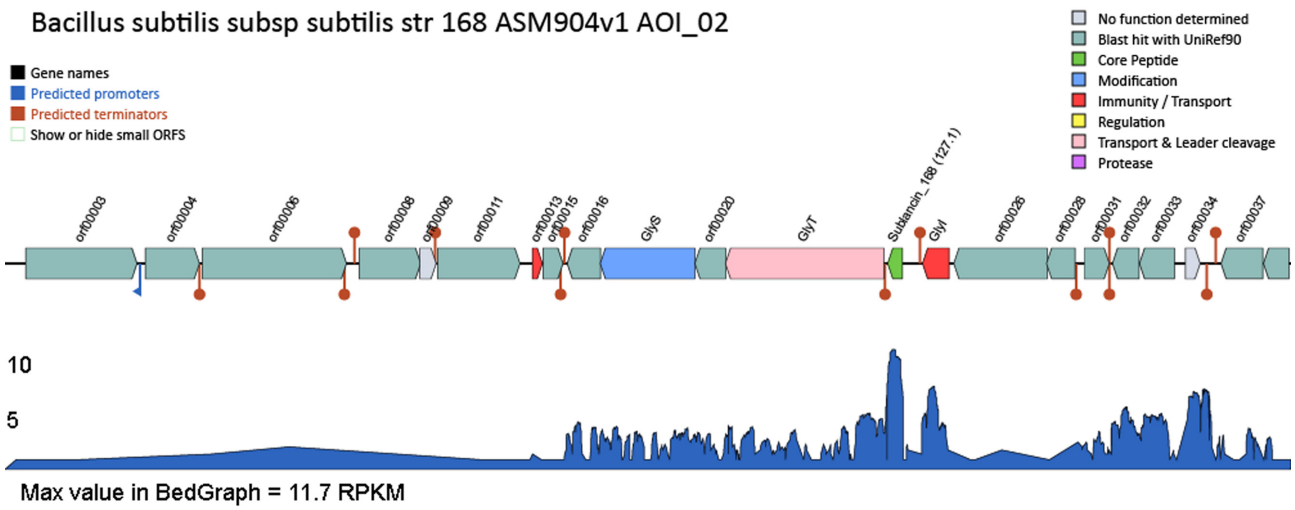

Figure 2. Example graphics produced by BAGEL4. On the left top items can be turned on or off by clicking on the item. Genes are indicated as arrows and additional information is displayed (including a link to BLAST the protein) by mouse-over. The additional information disappears by clicking on the gene. Bottom (in blue), RNA expression data displayed in RPMK.

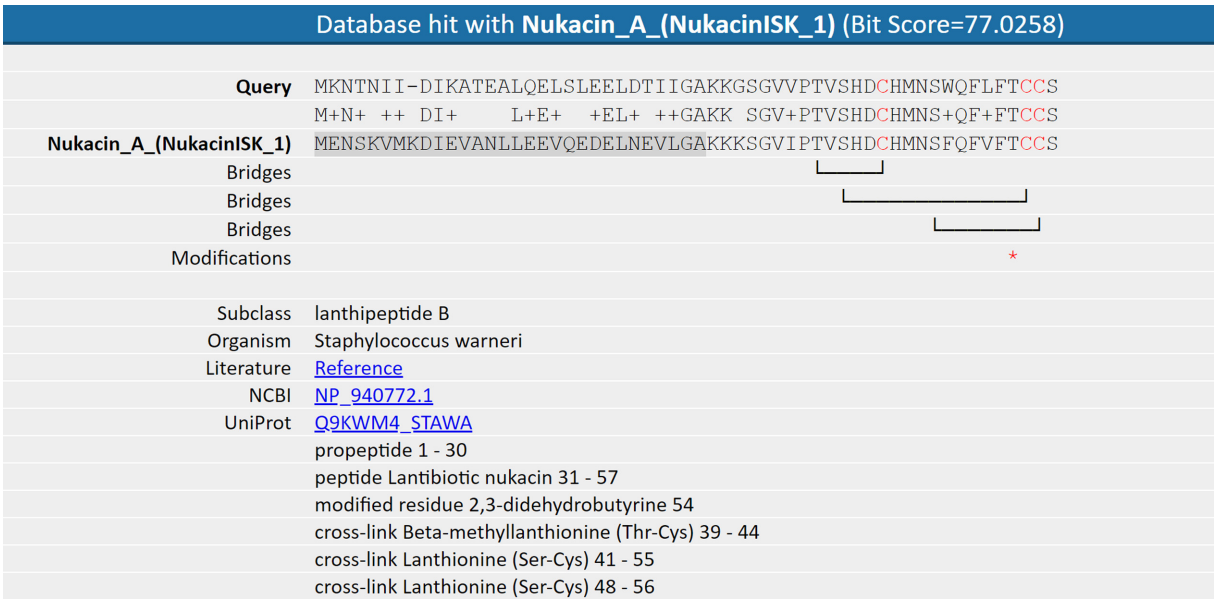

Figure 3. Example output of an alignment based on a BLAST hit with the core peptide database. The query is linked to a record (Nukacin A) in the core peptide database. Based on the UniProt identifier of the hit in the database, information available on modifications and bridging patterns is displayed. 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 σ^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

1. Arnison,P.G., Bibb,M.J., Bierbaum,G., Bowers,A.A., Bugni,T.S., Bulaj,G., Camarero,J.A., Campopiano,D.J., Challis,G.L., Clardy,J. *et al.* (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.*, **30**, 108–160.
2. Alt,S. and Wilkinson,B. (2015) Biosynthesis of the novel macrolide antibiotic anthracimycin. *ACS Chem. Biol.*, **10**, 2468–2479.
3. van Heel,A.J., Montalban-Lopez,M., Oliveau,Q. and Kuipers,O.P. (2017) Genome-guided identification of novel head-to-tail cyclized antimicrobial peptides, exemplified by the discovery of pumilarin. *Microb Genom*, **3**, e000134.
4. van Heel,A.J., Kloosterman,T.G., Montalban-Lopez,M., Deng,J.-J., Plat,A., Baudu,B., Hendriks,D., Moll,G.N. and Kuipers,O.P. (2016) Discovery, production and modification of 5 novel lantibiotics using the promiscuous nisin modification machinery. *ACS Synth. Biol.*, **5**, 1146–1154.
5. Qi,Y., D'Alessandro,J.M. and Blodgett,J.A.V. (2018) Draft genome sequence of streptomyces sp. Strain JV178, a producer of Clifednamide-Type polycyclic tetramate macrolactams. *Genome Announc.*, **6**, e01401–e01417.
6. Gerst,M.M., Dudley,E.G., Xiaoli,L. and Yousef,A.E. (2017) Draft genome sequence of bacillus velezensis GF610, a Producer of potent Anti-Listeria agents. *Genome Announc.*, **5**, e01046–e01017.
7. Borrero,J., Kelly,E., O'Connor,P.M., Kelleher,P., Scully,C., Cotter,P.D., Mahony,J. and van Sinderen,D. (2018) Plantaricyclin A, a novel circular bacteriocin produced by lactobacillus plantarum NI326: Purification, characterization, and heterologous production. *Appl. Environ. Microbiol.*, **84**, e01801–e01817.
8. de,Jong,A., van,Hijum,S.A.F.T., Bijlsma,J.J.E., Kok,J. and Kuipers,O.P. (2006) BAGEL: a web-based bacteriocin genome mining tool. *Nucleic Acids Res.*, **34**, W273–W279.
9. Blin,K., Wolf,T., Chevrette,M.G., Lu,X., Schwalen,C.J., Kautsar,S.A., Suarez Duran,H.G., de los Santos,E.L.C., Kim,H.U., Nave,M. *et al.* (2017) antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.*, **45**, W36–W41.
10. Skinnider,M.A., Merwin,N.J., Johnston,C.W. and Magarvey,N.A. (2017) PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.*, **45**, W49–W54.
11. Agrawal,P., Khater,S., Gupta,M., Sain,N. and Mohanty,D. (2017) RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. *Nucleic Acids Res.*, **45**, W80–W88.
12. Hammami,R., Zouhir,A., Le Lay,C., Ben Hamida,J. and Fliss,I. (2010) BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol.*, **10**, 22.
13. Waghu,F.H., Barai,R.S., Gurung,P. and Idicula-Thomas,S. (2015) CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res.*, **44**, D1094–D1097.
14. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., de Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C. *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
15. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.
16. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
17. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
18. Kingsford,C.L., Ayanbule,K. and Salzberg,S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol.*, **8**, R22.
19. van Heel,A.J., de Jong,A., Montalbán-López,M., Kok,J. and Kuipers,O.P. (2013) BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. *Nucleic Acids Res.*, **41**, W448–W453.