BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides

Auke J. van Heel[1,2], Anne de Jong[1,*], Manuel Montalbán-López[1], Jan Kok[1] and Oscar P. Kuipers[1,2,*]

[1]Molecular Genetics, University of Groningen, Linnaeusborgh, Nijenborgh 7, 9747AG Groningen, The Netherlands and [2]Kluyver Center for Genomics of Industrial Fermentation, Groningen/Delft, The Netherlands

ABSTRACT

Identifying genes encoding bacteriocins and ribosomally synthesized and posttranslationally modified peptides (RiPPs) can be a challenging task. Especially those peptides that do not have strong homology to previously identified peptides can easily be overlooked. Extensive use of BAGEL2 and user feedback has led us to develop BAGEL3. BAGEL3 features genome mining of prokaryotes, which is largely independent of open reading frame (ORF) predictions and has been extended to cover more (novel) classes of posttranslationally modified peptides. BAGEL3 uses an identification approach that combines direct mining for the gene and indirect mining via context genes. Especially for heavily modified peptides like lanthipeptides, sactipeptides, glycocins and others, this genetic context harbors valuable information that is used for mining purposes. The bacteriocin and context protein databases have been updated and it is now easy for users to submit novel bacteriocins or RiPPs. The output has been simplified to allow user-friendly analysis of the results, in particular for large (meta-genomic) datasets. The genetic context of identified candidate genes is fully annotated. As input, BAGEL3 uses FASTA DNA sequences or folders containing multiple FASTA formatted files. BAGEL3 is freely accessible at http://bagel.molgenrug.nl.

INTRODUCTION

Scientific interest in bacterial antimicrobial peptides and other posttranslationally modified peptides is increasing (1,2). Finding new antibiotic compounds from novel sources to fight multi-drug resistant pathogens has become the focus of many researchers. Furthermore, knowledge about the diverse enzymes involved in posttranslational modifications is rapidly advancing (3–5) and can be used to make new-to-nature antimicrobial peptides (6,7) or to stabilize medically relevant peptides (8). The discovered world of ribosomally synthesized and posttranslationally modified peptides (RiPPs) is constantly expanding. More and more modifications and the enzymes involved are being described (1). With the discovery of each new class new genome mining efforts are triggered. These efforts have led to valuable information and several high-impact publications (4,9–12). The main challenge in these kinds of genome mining efforts is the small size of the genes encoding the peptides of interest. Small open reading frames (ORFs) are often omitted during automated annotation efforts especially when their product sequences do not show strong homology with those of already described peptides, hampering a direct mining approach. Therefore, the large modification enzymes have been used regularly in indirect genome mining efforts. With the design and development of the BActeriocin GEnome mining tooL (BAGEL) since 2005, we aim to facilitate these efforts (13,14). Other useful tools have also been developed, such as the data repository Bactibase (15) and the prediction tool antiSMASH (16), which also supports non-ribosomal peptides but lacks some of the classes supported by the faster BAGEL3. In the current version of BAGEL, BAGEL3, our main goals were to combine direct and indirect mining, generate a

simpler, clearer and better quality output, make the analysis more independent of ORF predictions and to facilitate the addition of novel classes of peptides that can be mined for.

IMPLEMENTATION

New in BAGEL3

The major improvement in BAGEL3 is the new dual process (Figure 1), i.e. combining two mining strategies in one procedure. Another major advantage of BAGEL3 is its use of DNA sequences as input instead of annotated genomes, making it less dependent on ORF predictions. Furthermore, novel classes of RiPPs have been implemented, extending the genome mining capabilities of BAGEL3 beyond bacteriocins only. For this purpose new hidden Markov (HMM) models have been added describing specific genes involved in the biosynthesis of cyanobactins (called CyaG after PatG) (17), sactipeptides (called SacCD after TrnCD) (18) and linaridins (called LinL after CypL) (10).

BAGEL3 databases

BAGEL3 uses three different databases containing modified or unmodified bacteriocins and other posttranslationally modified peptides (non-bactericidal). The databases have been thoroughly updated. Each database contains all the records belonging to one of the three classes of proteins internal to BAGEL3: Class I contains posttranslationally modified peptides <10 kDa, the modification enzymes of which are encoded in the genomic context of the modified peptide and have been described for more than one case; Class II contains posttranslationally modified peptides <10 kDa not fitting the criteria of the first database; Class III contains anti-microbial proteins >10 kDa. This division is based on the procedure used by BAGEL3 to identify these proteins. The databases can be viewed online (http://bagel.molgenrug.nl/index. php/bacteriocin-database/) and have web links to literature, UniprotKB and NCBI. Users are actively encouraged to add new records to these databases via a web form (http://bagel.molgenrug.nl/index.php/submit-a-bacteriocin).

Description of the software

BAGEL3 uses DNA nucleotide sequences in FASTA format as input; multiple sequence entries per file are allowed. These DNA sequences are analyzed in parallel using two different approaches, one based on finding genes commonly found in the context of bacteriocin or RiPP genes, the other based on finding the gene itself.

The indirect approach (left red box in Figure 1) starts with performing a simple ORF call on the DNA. This call looks for ORFs of a certain minimal length that have a start and a stop codon not taking into account the presence of a possible ribosome-binding site. The products of these ORFs are subsequently screened for the presence of protein domains. Simple and defined rules based on these protein domains are then used to decide which part of the nucleotide sequence should be analyzed in more detail. These DNA sequences are called area(s) of interest (AOI). The size of the area is set to 20 K base pairs centered on the identified context gene. The ORFs in the AOI are then called using Glimmer (19). The next essential step is an additional specialized simple ORF call for every AOI to find the small ORFs that encode the targets of the identified modification enzymes. This ORF call takes into account the rule that was used to identify this AOI, so that when BAGEL3 is, for example, looking for a lanthipeptide, it will only call small ORFs encoding cysteine-containing peptides. Next, the context is annotated using the PFAM database (blast against Uniprot database is also possible in the stand-alone version). The last step is to identify the RiPP gene(s) that are present in the AOI. This is done using the results of both a Blast search against the BAGEL3 Class I and Class II databases and a screening for known motifs. If no direct hit is obtained then BAGEL3 predicts a precursor peptide sequence based on sequence properties and genomic organization.

The direct approach (right blue in box Figure 1) starts with a Glimmer ORF call. Next, the ORFs are blasted against all the three databases. The context (20 K base pair) of Blast hits is annotated using the PFAM database (a blast against Uniprot database is also possible in the stand-alone version).
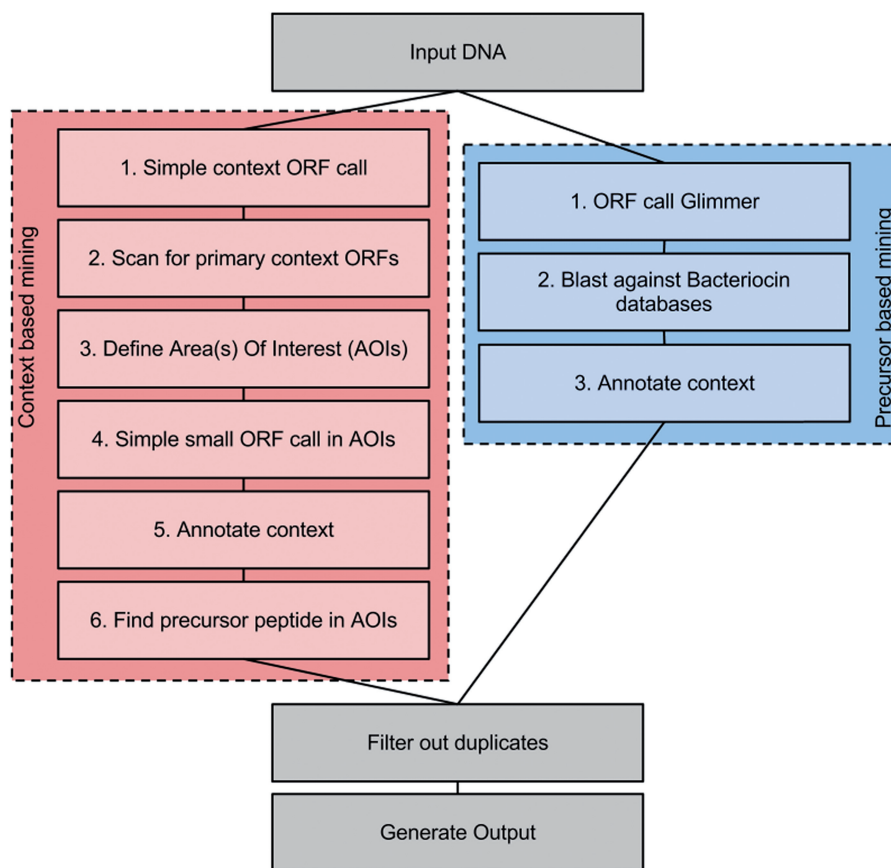
Because the same peptide could be identified with both approaches, the results of both are compared and filtered to exclude duplicates. Peptides identified via context genes are classified using this information (Table 1). Unmodified peptides identified via homology are classified according to their best Blast hit. Finally, an html output with graphics is generated from the large basic results table (see Figure 2). The whole process is logged into a log file. The nucleotide sequences of the identified AOIs can be downloaded.

Availability

The BAGEL3 web server can be accessed and used freely for files up to 20 MB (http://bagel.molgenrug.nl). A stand-alone Linux version is available on request for local installation. The stand-alone version can be easily adapted to personal preferences using a comprehensive configuration file.

System requirements

BAGEL3 runs on an Ubuntu Linux platform (http://www.ubuntu.com) with Apache web server (version2.2), MySQL server (version 14.14), PHP 5.4 (http://www.php.net/), Perl 5.10 (http://www.perl.org/), BioPerl 1.6.9 (http://www.bioperl.org/) and Joomla. Furthermore, the following software packages are used: BLAST 2.2.27 (20); HMMsearch 3.0 (http://hmmer.janelia.org/); Glimmer 3.02 (http://www.cbcb.umd.edu/software/glimmer/) (19), Pfamscan of the Sanger institute (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/) and the UniRef50 database (http://www.ebi.ac.uk/uniref/).

Figure 1. Schematic overview of the BAGEL3 genome mining procedure. BAGEL3 uses two different approaches in parallel to find bacteriocins and modified peptides. Both approaches use nucleotide sequences in FASTA format as input. The first approach (left, red) describes how the context-based approach proceeds. The second approach (right, blue) describes the simpler precursor peptide-based mining. Finally, both methods generate a single summary table with links to detailed graphical reports.

Table 1. Currently supported classes of RiPPs and the rules used to identify potential clusters

Name	Rule
Bottromycin	(PF04055) AND (PF02624)
Cyanobactin	(CyaG)
Glycocin	(TIGR04195) AND (PF03412)
Lanthipeptide class II	(PF05147) AND (PF13575)
Lanthipeptide class I	(PF04737|PF04738|PF14028) AND (PF05147)
Lanthipeptide class III	(lanKC)
Lanthipeptide class IV	(PF05147) AND (LanL)
LAPs	(PF02624) AND (PF00881)
Lasso peptide	(PF13471) AND (PF00733)
Linaridin	(LinL)
Microcin	(PF02794)
Sactipeptides	(SacCD) AND (PF04055)
Thiopeptide	(PF02624) AND (PF00881) AND (PF14028)

\| = or AND = additional requirement. The rules in this table describe the criteria that have to be matched by a certain stretch of DNA to become an AOI. Some rules might overlap, and therefore they are checked in an ordered fashion. In this way, the more stringent rule is checked after the less stringent.

Validation of BAGEL 3

The BAGEL3 software was validated using a set of 50 genomes known to encode bacteriocins and other modified peptides. It was checked that no known compounds were missed. Next, to validate if novel clusters could be identified, 200 draft genome sequences from the NCBI server were screened. Both these sets of genomes were also used to check for false positives. A false positive was defined as a cluster that did not have at least a likely core peptide or a gene context that can be associated with RiPP biosynthesis.

RESULTS AND DISCUSSION

Analysis of example genomes

Based on the newly added HMM CyaG, which identifies the serine protease, generally termed G protein, in the cyanobactin biosynthesis pathway (21), we found a new cyanobactin encoded in the genome of the cyanobacterium *Lyngbya* sp PCC8106 (see Table 2). In *Enterococcus faecalis* Fly1, BAGEL3 identified an interesting novel lantibiotic gene cluster. The cluster could code for two lanthipeptides, which is common for two-component
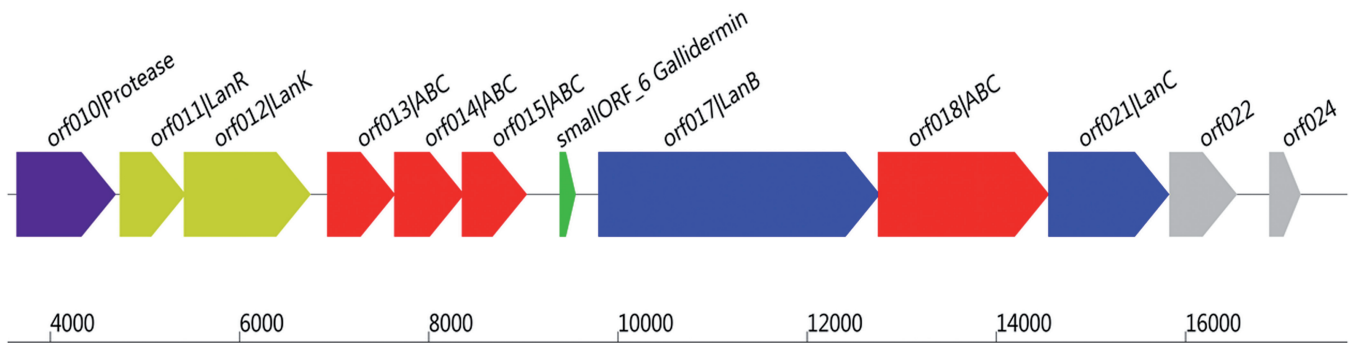
Figure 2. Example detailed report of a lantibiotic cluster encoding a nisin variant and its modification enzymes found in *Streptococcus suis* J14 (NC_017618.1) using BAGEL3. The target gene (smallORF_6) was in this case identified by the specialized small ORF calling procedure.

lantibiotics, but in this case they are not modified by a LanM type enzyme but by a single set of LanBC enzymes. Another example of the added value of BAGEL3 is demonstrated when querying the plasmid pTEF2 of *E. faecalis* V538. Based on the context, BAGEL3 identifies a glycocin-like peptide (Table 2). Glycocins are glycosylated antimicrobial peptides of which Glycocin S and Sublancin 168 are the only two characterized members that also contain disulfide bridges (1). The identified peptide has an exact match with the

Table 2. A selection of novel RiPPs identified by BAGEL3

DNA screened	Homology (*P*-value)	Identification	Sequence
P. larvae subsp *larvae* BRL 230010 Ctg01135	Sporulation-killingfactor_skfA [1e-10]	Context: SacCD	Sactipeptide: MSNHNVRNEPAPAWESSAQNNLSKPAGIPLIKSVGCAACWGAK NISLTRACLPPTPINLAL
pTEF2 *E. fecalis* V538	Enterocin_96 [2e-41] (exact match)	context: TIGR04195 PF03412	Glycocin: MLNKKLLENGVVNAVTIDELDAQFGGMSKRDCNL MKACCAGQAVTYAIHSLLNRLGGDSSDPAGCNDIVRKYCK
E. faecalis Fly1 cont1.76	leader_abc mature_ab PF02052.7 leaderLanBC	context: PF04737.5 PF04738.5 PF05147.5	Lantibiotic: MPKYDDFDLNLKQTSASNQKDTRVTSVMACTPGTCNNKCPN TNWLCSNVCVTK1CWTCA
E. faecalis Fly1 cont1.76	leader_abc PF02052.7 leaderLanBC	context: PF04737.5 PF04738.5 PF05147.5	Lantibiotic: MPKYDDFDLNLKQNVSSSNKEPRITSIKWCTPG TCNNTCKGDSTLKSNCCGGSLMCSLGGC
Lyngbya sp PCC 8106	Trunkamide[1e-06]	Context: CyaG	Cyanobactin: MPCYPSYDGVDASVCMPCYPSYDGVDASVCMPCYP SYDDAE
B. pseudomallei 354a Contig0218	Capistruin[5e-23]	Context: PF13471.1 PF00733.16	Lasso peptide: MVRFLAKLLRSTIHGSHGVSLDAVSSTHGTPGFQTPDARV ISRFGFN

previously described bacteriocin Enterocin 96 (22). In the article describing Enterocin 96, the authors note that the measured mass is higher than the theoretical peptide mass. This mass difference is perfectly in line with the BAGEL3 predicted glycosylation, which has also been suggested by others (23). In the genome of the honey bee pathogen *Paenibacillus larvae* subsp *larvae* BRL 230010, BAGEL3 identified the gene for a so-called sactipeptide that shows low homology to sporulation killing factor SfkA of *Bacillus subtilis* 168. In the genome of the gram negative pathogen *Burkholderia pseudomallei*, 354a, a lasso peptide, was identified with strong homology to capistruin (24).

Simple small ORF calling procedure facilitates genome mining

A big problem when screening large amounts of genomic data using BAGEL2 was its dependence on the quality of the annotation. Multiple different ORF prediction procedures were therefore implemented. Consequently, several results had to be compared while still some of the small ORFs encoding antimicrobial peptides were lacking, creating the need for manual evaluation of some of the identified gene clusters. To remove this dependency, we now use a simple small ORF calling procedure for the AOIs identified by their context, obviating the need for reannotation and simplifying the procedure and the analyses.

BAGEL3 is extensible

The novel mining approach used in BAGEL3 has the big advantage that it can easily be extended to include new classes of modified peptides, the only requirement being that the gene of the small peptide of interest lies in a genomic context that can be recognized. The genomic context has to be translated into a simple rule that describes a certain AOI (for examples see Table 1). The described precursor peptides should then be added to the database. If requirements for the precursor peptides are known (for example, must contain a cysteine), these can be added. The context rule should be tested to check if it is specific enough. Adding a new class of peptides to the system can be done in within a few hours if an extensive literature review is available. Users are encouraged to submit novel rules/classes via the online form available on the BAGEL3 web site.

CONCLUSIONS

BAGEL3 is a versatile fast genome-mining tool valid not only for modified and non-modified bacteriocins but also for non-bactericidal RiPPs. It can handle large data sets like those from metagenome projects. This updated version looks for bacteriocins/RiPPs via two different approaches, which increases the success rate and lowers the need for manual evaluation of results. The new design also allows for easy inclusion of novel classes of peptides, and users are therefore encouraged to propose the addition of novel classes.

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. Cotter,P.D., Ross,R.P. and Hill,C. (2012) Bacteriocins—a viable alternative to antibiotics? *Nat. Rev. Microbiol.*, **11**, 95–105.
3. Zhang,Q., Yu,Y., Vélasquez,J.E. and van der Donk,W.A. (2012) Evolution of lanthipeptide synthetases. *Proc. Natl Acad. Sci. USA*, **109**, 18361–18366.
4. Freeman,M.F., Gurgui,C., Helf,M.J., Morinaka,B.I., Uria,A.R., Oldham,N.J., Sahl,H., Matsunaga,S. and Piel,J. (2012) Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science*, **338**, 387–390.
5. Oman,T.J., Boettcher,J.M., Wang,H., Okalibe,X.N. and van der Donk,W.A. (2011) Sublancin is not a lantibiotic but an S-linked glycopeptide. *Nat. Chem. Biol.*, **7**, 78–80.
6. Huo,L., Rachid,S., Stadler,M., Wenzel,S.C. and Müller,R. (2012) Synthetic biotechnology to study and engineer ribosomal bottromycin biosynthesis. *Chem. Biol.*, **19**, 1278–1287.
7. van Heel,A.J., Mu,D., Montalbán-López,M., Hendriks,D. and Kuipers,O.P. (2013) Designing and producing modified, new-to-nature, peptides with antimicrobial activity by use of a combination of various lantibiotic modification enzymes. *ACS Synth. Biol.*, doi: 10.1021/sb3001084.
8. Kluskens,L.D., Nelemans,S.A., Rink,R., de Vries,L., Meter-Arkema,A., Wang,Y., Walther,T., Kuipers,A., Moll,G.N. and Haas,M. (2009) Angiotensin-(1-7) with thioether bridge: an angiotensin-converting enzyme-resistant, potent angiotensin-(1-7) analog. *J. Pharmacol. Exp. Ther.*, **328**, 849–854.
9. Maksimov,M.O., Pelczer,I. and Link,A.J. (2012) Precursor-centric genome-mining approach for lasso peptide discovery. *Proc. Natl Acad. Sci. USA*, **109**, 15223–15228.
10. Claesen,J. and Bibb,M. (2010) Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proc. Natl Acad. Sci. USA*, **107**, 16297–16302.
11. Velásquez,J.E. and Van Der Donk,W.A. (2011) Genome mining for ribosomally synthesized natural products. *Curr. Opin. Chem. Biol.*, **15**, 11–21.
12. Lee,S.W., Mitchell,D.A., Markley,A.L., Hensler,M.E., Gonzalez,D., Wohlrab,A., Dorrestein,P.C., Nizet,V. and Dixon,J.E. (2008) Discovery of a widely distributed toxin biosynthetic gene cluster. *Proc. Natl Acad. Sci. USA*, **105**, 5879–5884.
13. 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.
14. de Jong,A., van Heel,A.J., Kok,J. and Kuipers,O.P. (2010) BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res.*, **38**, W647–W651.
15. Hammami,R., Zouhir,A., Le Lay,C., Hamida,J.B. and Fliss,I. (2010) BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol.*, **10**, 22.
16. Medema,M.H., Blin,K., Cimermancic,P., de Jager,V., Zakrzewski,P., Fischbach,M.A., Weber,T., Takano,E. and Breitling,R. (2011) antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, **39**, W339–W346.
17. Schmidt,E.W., Nelson,J.T., Rasko,D.A., Sudek,S., Eisen,J.A., Haygood,M.G. and Ravel,J. (2005) Patellamide A and C biosynthesis by a microcin-like pathway in prochloron didemni, the cyanobacterial symbiont of lissoclinum patella. *Proc. Natl Acad. Sci. USA*, **102**, 7315–7320.
18. Rea,M.C., Sit,C.S., Clayton,E., O'Connor,P.M., Whittal,R.M., Zheng,J., Vederas,J.C., Ross,R.P. and Hill,C. (2010) Thuricin CD, a posttranslationally modified bacteriocin with a narrow spectrum of activity against clostridium difficile. *Proc. Natl Acad. Sci. USA*, **107**, 9352–9357.
19. 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.
20. 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.
21. Schmidt,E.W., Nelson,J.T., Rasko,D.A., Sudek,S., Eisen,J.A., Haygood,M.G. and Ravel,J. (2005) Patellamide A and C biosynthesis by a microcin-like pathway in prochloron didemni, the cyanobacterial symbiont of lissoclinum patella. *Proc. Natl Acad. Sci. USA*, **102**, 7315–7320.
22. Izquierdo,E., Wagner,C., Marchioni,E., Aoude-Werner,D. and Ennahar,S. (2009) Enterocin 96, a novel class II bacteriocin produced by enterococcus faecalis WHE 96, isolated from munster cheese. *Appl. Environ. Microbiol.*, **75**, 4273–4276.
23. Stepper,J., Shastri,S., Loo,T.S., Preston,J.C., Novak,P., Man,P., Moore,C.H., Havlcek,V., Patchett,M.L. and Norris,G.E. (2011) Cysteine S-glycosylation, a new post-translational modification found in glycopeptide bacteriocins. *FEBS Lett.*, **585**, 645–650.
24. Knappe,T.A., Linne,U., Zirah,S., Rebuffat,S., Xie,X. and Marahiel,M.A. (2008) Isolation and structural characterization of capistruin, a lasso peptide predicted from the genome sequence of burkholderia thailandensis E264. *J. Am. Chem. Soc.*, **130**, 11446–11454.