BAGEL: a web-based bacteriocin genome mining tool

Anne de Jong, Sacha A. F. T. van Hijum*, Jetta J. E. Bijlsma, Jan Kok and Oscar P. Kuipers

Department of Molecular Genetics, University of Groningen, Groningen Biomolecular Sciences and Biotechnology Institute, PO Box 14, 9750 AA Haren, The Netherlands

ABSTRACT

A common problem in the annotation of open reading frames (ORFs) is the identification of genes that are functionally similar but have limited or no sequence homology. This is particularly the case for bacteriocins, a very diverse group of antimicrobial peptides produced by bacteria and usually encoded by small, poorly conserved ORFs. ORFs surrounding bacteriocin genes are often biosynthetic genes. This information can be used to locate putative structural bacteriocin genes. Here, we describe BAGEL, a web server that identifies putative bacteriocin ORFs in a DNA sequence using novel, knowledge-based bacteriocin databases and motif databases. Many bacteriocins are encoded by small genes that are often omitted in the annotation process of bacterial genomes. Thus, we have implemented ORF detection using a number of published ORF prediction tools. In addition, BAGEL takes into account the genomic context, i.e. for each potential bacteriocin-encoding ORF, the sequence of the surrounding region on the genome is analyzed for genes that might encode proteins involved in biosynthesis, transport, regulation and/or immunity. These innovations make BAGEL unique in its ability to detect putative bacteriocin gene clusters in (new) bacterial genomes. BAGEL is freely accessible at: http:// bioinformatics.biol.rug.nl/websoftware/bagel.

INTRODUCTION

Bacteriocins are antimicrobial peptides produced by bacteria, which are active against either closely related or more distant species. They provide a defense mechanism for the producing strain as they can kill other bacteria. Therefore, bacteriocins are applied as food preservatives (1,2) and are of interest for the development of novel antibiotics (3,4). They are exported across the cytoplasmic membrane by dedicated transporters containing an ATP-binding cassette (ABC-transporter), and are often processed by a specific protease, although occasionally these two functions are combined (5). In many cases the bacteriocin-encoding gene cluster also contains one or more immunity proteins to prevent self-killing. The expression of bacteriocin gene clusters is often under control of a two-component signal transduction system, which is usually part of the cluster. The inducer can be either the bacteriocin itself or a bacteriocin-like peptide. Various classes of over 200 known bacteriocins have been defined, based on features such as the nature of post-translational modifications, specific anti-bacterial activity, formation of oligomers, protein size, presence of sugar moieties, presence of positively charged amino acids and mode of action. Five main classes reported in literature are as follows: (i) lantibiotics, posttranslationally modified peptides (5); (ii) non-modified heat stable bacteriocins (6–8); (iii) large heat-labile bacteriocins; (iv) complex bacteriocins carrying lipid or carbohydrate moieties (9) and (v) circular bacteriocins (10) (Table 1). A number of classes are divided into subclasses (Table 1). These differences in properties reflect the large divergence between bacteriocins (11).

The classical way of identification of a bacteriocin has been by determining its biological activity through extensive testing of the (putative) producer strain for inhibition of the growth of other bacteria. A few reports (5,7,9,12) describe the identification of putative bacteriocins by screening a genomic DNA sequence for the presence of bacteriocin genes and their genomic context for biosynthetic genes.

Here, we present the web-based software tool BAGEL, which enables the identification of bacteriocins and their biosynthetic clusters through a knowledge-based database. It takes advantage of the fact that accessory genes encoding proteins needed for processing, modification, transport, regulation and/or immunity are commonly located in the vicinity of a putative bacteriocin gene. Furthermore, open reading frame (ORF) detection is provided, which makes BAGEL independent of GenBank annotations and thus prevents the oversight of small non-conserved ORFs (the most probable candidates for bacteriocin genes), which are omitted from many genome annotations. A typical BAGEL search on a genome sequence results in a set of putative bacteriocin gene clusters. These are ranked according to the presence of significant features in the amino acid sequences and their genomic context. The output contains comprehensive information on the predicted putative bacteriocins. BAGEL is the first fully automated and very fast tool for the

*To whom correspondence should be addressed. Tel: +31 50 3632092; Fax: +31 50 3632348. Email: S.A.F.T.van.Hijum@rug.nl

Table 1. Overview of the known classes of bacteriocins and the features that discriminate the different classes

Class	Type	Special features	Mature size in kDa	pI	PFAM domain	Charge	Processing site	Motifs	Adjacent genes
IA	Lantibiotics	Posttranslationally modified Elongated shaped molecules	<5	>8	PF05500 PF04369		G(SA) P(RQ)	F(ND)L(DEN)(LVI) SLCTPGC SXXXCPTTXCXXXC	ABC-transporter His-kinase C39-protease
IB	Lantibiotics	Posttranslationally modified Globular molecules	<5	>8	PF05500 PF04369		G(SA) P(RQ)	F(ND)L(DEN)(LVI) FTCCS GXXXTOBX-C	ABC-transporter His-kinase
IIa	Non-lantibiotics	Strongly cationic Heat stable Strong anti-lysterial activity Pediocin-like at least one disulfide bridge	<10	>9	PF04604 PF02052 PF01721	Strongly Cationic	GG	YGNGVXC LSXXELXXIIXGG DoubleGG	ABC-transporter His-kinase C39-protease
IIb	Two-peptide non-lantibiotics	Possess at most one positive charge	25–65	>9	PF02052 PF01721	Weak	GG P(RQ)	GXXXTOBX-C LSXXELXXIIXGG DoubleGG	ABC- transporter His-kinase C39-protease
IIc		No leader	30–65	?		Cationic			ABC-transporter His-kinase
II microcin				>7			G(GSA)	MREOB	ABC-transporter His-kinase C39-protease
III	Large heat labile	Large	>30	?					
IV		Complex bacteriocins carrying lipid or carbohydrate moieties	>10	?					
V	Circular	Two transmembrane segments	49–108	9.5–11					CirC

identification of new bacteriocin gene clusters. We demonstrate the power and versatility of this software by the analysis of a number of annotated and non-annotated bacterial genomes.

IMPLEMENTATION

System requirements and the web-interface

BAGEL runs on a Linux platform (Fedora Core 3; http://fedora.redhat.com/) with Apache web-server (2.0.48), MySQL server (version 3.23.58), PHP 4.3 (http://www.php.net/) and Perl 5.8.7 (http://www.perl.org/). Furthermore, the following software is used: FASTA 3.4 (13); Blast 2.2.9 (14); HMMsearch (HMMER 2.2g HHMI/Washington University School of Medicine); Glimmer v2.13 (15) / RBSfinder (http://www.tigr.org/software/genefinding.shtml); Zcurve (16); and GeneMark (17). Depending on the load of the server (in this study a dual Opteron 2.2 GHz was used) a BAGEL search takes about 1 min to discover bacteriocin gene clusters in a genome of 2 Mb. In short, BAGEL consists of a PHP web-interface and a bash script with three Pascal modules compiled by FreePascal (version 1.0.10; http://www.freepascal.org/). These modules allow connecting the various tools employed by BAGEL (see above) by their input and output.

The web-interface consists of three separate web pages: (i) one for uploading of a GenBank file; (ii) a page with parameters; and (iii) the status and results page. The status page presents, after a successful run, hyperlinks to the search results web pages. Each run is assigned a session-id allowing the user to inspect the results at any given time. The output of BAGEL consists of several graphical representations of the genomic context of the putative bacteriocin encoding gene (5).

Process overview

Figure 1 presents a flow scheme of BAGEL. The first step (Figure 1A) provides a genome sequence in the form of a GenBank file. The second step (Figure 1B) enables to set parameters for the search. In the third step (Figure 1C), the run phase, the following is performed: (i) screening of the provided genome sequence for protein-databases (bacteriocin, highly unlikely bacteriocins, immunity and circularization proteins) and motif-databases; (ii) properties are added to the putative bacteriocins and (iii) the results are combined and exported to a result database. A successful run results in a web page (Figure 1D) with information concerning the putative bacteriocins.

Input

Genome sequences in GenBank format are used by BAGEL. They can either be selected from an extensive list of sequenced bacterial genomes (NCBI: ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) or be provided by the user. In the case that a non-annotated GenBank file is uploaded, a reference genome can optionally be selected from the available genomes. The use of a reference genome, which should be the annotated counterpart of the uploaded genome, allows the user to quickly determine whether or not a putative bacteriocin-encoding ORF has already been annotated in
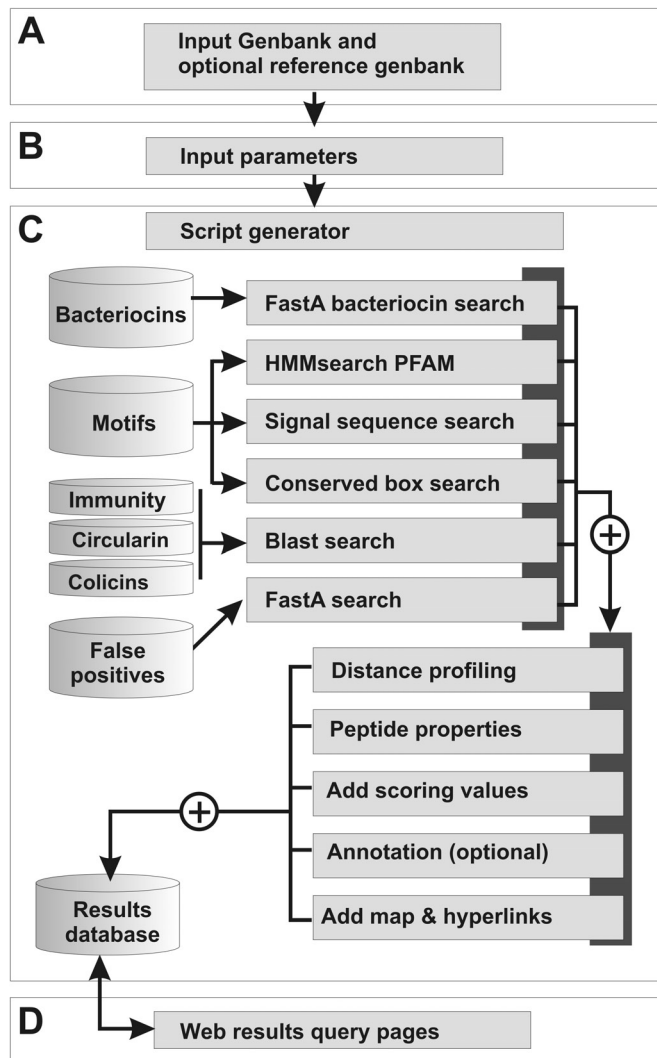

Figure 1. Schematic overview of BAGEL. (**A**) A web page where a genome sequence is provided in GenBank file format. Optionally, an annotated reference genome can be selected in case a *de novo* ORF search was performed on the input genome. (**B**) A web page with parameters concerning the search and scoring of putative bacteriocins. (**C**) The actual search for putative bacteriocins takes place through a script which recruits (i) the various search modules; (ii) the peptides with a hit to the various databases (indicated by a plus sign) are further annotated; and (iii) the results concerning putative bacteriocins (indicated by a plus sign) are exported to a results database. (**D**) This database is used to provide the BAGEL user with the search results in a web page. Annotation (optional): if a genome with new ORF annotations was provided, the (optional) annotated reference genome is used (if possible) to connect the putative bacteriocin ORF to an existing genome annotation.

the genome of interest. In addition, we offer a freeware tool, Genome2D (18), to convert several formats (FASTA, tab delimited or Excel files) to the GenBank file format.

ORF detection

Because bacteriocins are commonly encoded by small ORFs that are regularly omitted in the annotation process of bacterial genomes, we provide ORF detection. In order to deal with genome sequences of different GC contents, three ORF prediction tools suited for detection of small ORFs

have been implemented in our ORF tool: (i) Glimmer/RBSfinder (15); (ii) Zcurve (16) and (iii) GeneMark (17). In a DNA sequence in FASTA file format, the ORFs are detected and saved in a GenBank formatted file that can subsequently be used for bacteriocin detection by BAGEL.

Bacteriocin database

A bacteriocin database was built to enable comparison of peptides being screened to known bacteriocins and their gene clusters. Bacteriocin sequences were retrieved from various databases: (i) the SRS server of ExPasY (http://www.expasy.org/srs/); (ii) the NCBI server (http://www.ncbi.nlm.nih.gov/entrez/); and (iii) the text search option from UniProt (http://www.expasy.uniprot.org/). Since not all known bacteriocins were present in these databases, expertise of our research group in combination with a literature search was used to complete the bacteriocin database. The annotation of these bacteriocins was extended with information of bacteriocin classes and, if possible, a hyperlink to the UniProt database. A database of known colicins, derived from the NCBI database, was built to screen with Blast for colicins. These class-III bacteriocins are identified solely on the basis of mutual homology because they are relatively large and well conserved.

Motif database

A number of conserved short peptides (motifs) have been described for specific bacteriocins such as the FNDLV motif and GG motif (7,19). These motifs enable the search in poorly conserved proteins. We extended the number of motifs from conserved regions in bacteriocins (Table 2) based on literature data. For example, the motif for a common

Table 2. The motifs described in literature that were included in the BAGEL motif- and leader- sequence-databases

Motif	Motif present in	Reference
YGNGXXCXXXXC	Class IIa	(25)
GXXXTXBEC	Class IIb	(26)
YYGNGVXC	Class IIa	(27)
GWAXGXXXG	Class II	(19)
FNDLV	Lantibiotics	(19)
MREOB	Class II microcins	(28)
GXOXTXBX-C	Lantibiotics	(26)
CPTTXCXXXC	Lantibiotics	(26)
AXXXAAXGA	Lantibiotics	This work
AX(FA)(LFV)AA(PL)GA	Lantibiotics	This work
Leader motifs		
PR/PQ/GA/GS	Lantibiotics and Class IIb	(20)
GG-motif	Class IIa and IIb	(7)

processing site contains two conserved glycine residues in positions −1 and −2 (double-GG motif) (7) relative to the start of the mature peptide. This double-GG motif was extended with GA, GS, PR, PQ (20). These two amino acid sites are only considered as potential processing sites if they are within a proper range from the N-terminus (by default 12–25 amino acids). From the alignment of the double-GG leader sequences of a number of bacteriocins (7), a weight matrix was constructed that scores for a processing site and the leader sequence.

The adjacent genes of bacteriocin-encoding genes contain well-defined conserved motifs, which are described in a PFAM database (in the form of hidden Markov models). The domains for ABC-transporter and the protease C39 family were described by the well-defined existing PFAM domains, PF00005 and PF03412, respectively (Table 1). PFAM domain PF06580 represents a conserved region within bacterial histidine kinase enzymes (Table 1).

Database with highly unlikely candidates (HUC-database)

Several genomes were screened for bacteriocins genes (see Supplementary Table S1) and manually investigated for the presence of highly unlikely candidates (HUCs) based on (i) their annotation in EMBL/GenBank database, (ii) similarity to other organisms, and (iii) expert knowledge. Those candidates that complied with the above-mentioned criteria were used for the HUC-database. BAGEL will consider a peptide as an extremely unlikely bacteriocin candidate if the FASTA score of the peptide to a member of the HUC-database is lower than 10^{-8}.

Classification

Five classes of bacteriocins have been described in literature (21) (Table 1). In order to determine the class to which a putative bacteriocin belongs, five criteria were used; (i) homology to a classified bacteriocin from the bacteriocin database; (ii) presence of class-specific PFAM domains (Table 1); (iii) presence of class-specific motifs from the motif database; (iv) protein size; and (v) iso-electric point of the mature peptide being eight or higher. Trial runs with known bacteriocins showed that, although in many cases the correct class was predicted, in some cases no class could be assigned, indicating that these criteria are not discriminative for all putative bacteriocins. Bacteriocins belonging to class III, consisting of relatively large (>30 kDa) bacteriocins also known as colicins, are easily predicted as they are very homologous to their known counterparts.
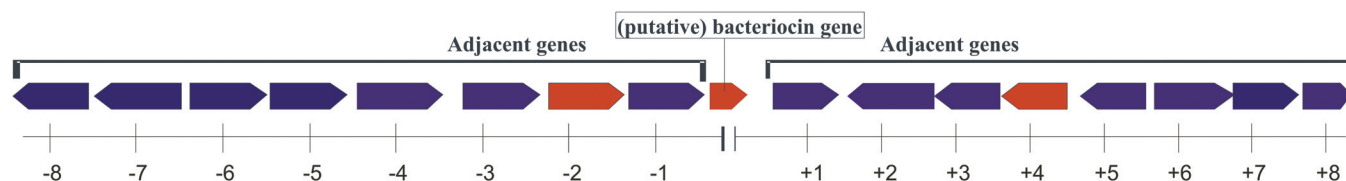

Figure 2. Distance profiling. Adjacent genes are annotated on the basis of motifs or a hit to one of the databases described in Figure 1C. The presence of ORFs, of which the products are homologous to proteins known to be involved in bacteriocin biosynthesis, contributes to the probability that an ORF encodes a bacteriocin. In this example, the biosynthetic genes (red) are at position −2 (two ORFs upstream) and +4 (four ORFs downstream) of the putative bacteriocin ORF, as indicated by the ruler.

Table 3. Properties used for the scoring of a putative bacteriocin gene

Property	Weight factor
Motif present	60
FASTA hit on bacteriocin database	100
Adjacent genes	100
PFAM/HMM motif	60
Leader sequence present	100
Blast hit with colicin database	120
Charge, pI, Cysteins.	5

The sum of the weight factors for the scored properties yields the final score. The peptide is considered as a significant hit when the total score is above 175.

Search criteria

Six steps have been implemented to search for putative bacteriocins: (i) a FASTA search for known bacteriocins in the bacteriocin database (default cut-off is 10^{-4}); (ii) BLAST search for known colicins in the colicin database (default cut-off is 10^{-15}); (iii) a HMMsearch with PFAM domains from the motif database (Table 1) (default cut-off is 10^{-1}); (iv) leader sequence detection by using the motif database (7); (v) conserved motif search with a regular expression (Table 2); and (vi) distance profiling (default distance is 8, which means that biosynthetic genes should be at a maximum distance of 8 ORFs to the putative bacteriocin) (Figure 2). An HMM-search is used to annotate biosynthetic genes (default cut-off is 10^{-4}). After a search, data on the protein size, charge distribution, pI and cystein content are added to the results. Not all of the mentioned criteria contribute equally to the final score. A hit with a search method results in the increase of the score of a putative bacteriocin by an associated weight factor (Table 3). These weight factors have been empirically fine tuned by screening 10 bacterial genomes (see Supplementary Table S1) to yield the best signal (an experimentally verified bacteriocin) to noise (a very unlikely candidate bacteriocin; see above) ratio. A putative bacteriocin is discarded by BAGEL if it (i) is similar to a member of the HUC-database (default Blast cut-off is 10^{-16}) or (ii) is smaller than the minimum size of a bacteriocin peptide (default 25 amino acids).

Output

After the screening of the protein sequences for putative bacteriocins, the results are exported to a MySQL database. The status- and results web page shows a table with a summary of the search results. In addition, two queries on the database can be performed by the user: (i) a list of results in a table including hyperlinks to web resources and information of significant properties and (ii) a table for each potential bacteriocin and a visual representation of its location on the genome sequence.

Availability

For performing a *de novo* search for bacteriocins the BAGEL web-server is freely accessible at http://bioinformatics.biol.rug.nl/websoftware/bagel. The bacteriocin database can be queried from the web interface at http://bioinformatics.biol.rug.nl/bacteriocin. The ORF prediction tool is freely accessible at http://bioinformatics.biol.rug.nl/websoftware/orf. Table S1 is accessible at http://bioinformatics.biol.rug.nl/supplementary/bagel_data/.

RESULTS AND DISCUSSION

Design and validation

Owing to the large diversity in bacteriocins (11) and their small size, a complex strategy is needed for the automated detection of their encoding genes in bacterial genomes. To this end, BAGEL was developed. Ideally, *de novo* ORF detection is performed on a genome sequence to identify all (small) ORFs. The resulting ORFs are screened by BAGEL using a reference genome to annotate the resulting putative bacteriocin genes.

The BAGEL screening process of ORF products (by default sizes between 25 and 100 amino acids) starts by a FASTA search (13) against our bacteriocin database. FASTA was found to be more accurate for small peptides than BLAST (14). Larger peptides (>100 amino acids) are more efficiently screened with the Blast alignment method. The putative bacteriocin sequence is annotated by comparison with the various databases (see above for details).

The methodology employed by BAGEL has been optimized empirically by a constant evaluation of the performance of the software in detecting known bacteriocins. For four bacterial genomes (Table 4 and Supplementary Table S1) BAGEL correctly identified all 19 encoded and described bacteriocins, demonstrating the high accuracy of the software.

Identification of new bacteriocins with BAGEL

In addition to sublancin and subtilisin, four other ORFs in the genome of *Bacillus subtilis* were identified as putative bacteriocin-encoding genes (Table 4). The peptides specified by these four ORFs contain a number of bacteriocin-like features (Table 4). Two peptides, YhaJ and YukD, show weak similarity to microcin H47. The other two peptides, *yxzE* and *yufS*, are not similar to any known bacteriocins. However, ORFs in the immediate vicinity of the *yxzE* and *yufS* genes specify proteins that are homologous to (i) an ABC transporter and (ii) a two-component system, demonstrating that the context search might lead to the discovery of new types of bacteriocin-like peptides.

In *Streptococcus pneumoniae* TIGR4 seven putative bacteriocin genes (*blpUKNIJMO*) have been described (22), which were all identified by BAGEL. Interestingly, of these seven only *blpU* (*thmA* in R6) (23,24) was identified by BAGEL in the R6 strain (Table 4). Four additional putative peptides were scored as potential bacteriocins encoded by the genome of *S.pneumoniae* TIGR4 (Table 4). They all are annotated as 'hypothetical' and do not show any homology to known bacteriocins. However, they all contain a potential processing site and their adjacent genes encode proteins that could putatively be involved in the production of an active bacteriocin, again demonstrating the power of the context search for the identification of novel putative bacteriocins or inducing factors. This BAGEL search indicates that *S.pneumoniae* TIGR4 could contain as many as

Table 4. Summary of results obtained by the application of BAGEL on various genomes

Organism	gene/locus	Homology	ABC	HK	C39	Score	Reference
Bacillus subtilis 168	*sunA*	Sublancin	−1		−1	375	(20)
	sbo	Subtilosin A				255	(12)
	yhaJ	Microcin H47				255	
	yukD	Microcin H47				255	
	yxzE		−5/+2			225	
	yufS		−4	−7		225	
Streptococcus	SP0041	BlpU	+1		+1	385	(22)
pneumoniae TIGR4	SP0533	BlpK				360	(22)
	SP0540	BlpN				250	(22)
	SP0531	BlpI				255	(22)
	SP0532	BlpJ				255	(22)
	SP0539	BlpM				255	(22)
	SP0541	BlpO				255	(22)
	SP0109	putative	+2			325	
	SP1949		+8		+4	175	
	SP0792		−6	+7		225	
	SP1832		−7/+8			165	
	SP1948		+5		+5	165	
Streptococcus	*thmA*	BlpU	3		3	490	(23,24)
pneumoniae R6	spr0141		−4/8			175	
	spr1185		−2/3			235	
	spr1199		−8/4			235	
	spr1766		7		4	175	
	ip		3		4	325	
	spr0700		−6	+8		225	
	spr0098	SP0109	8			325	
	spr1765		+8		+5	165	

In the columns ABC (ABC transporters), HK (Histidine kinase), C39 (protease), the distance in number of ORFs of these genes to the putative bacteriocin gene is indicated (Figure 2). All peptides derived from these putative bacteriocin genes contain a processing site.

11 bacteriocin-encoding sequences, but as no bacteriocin-like activity has been proven for this strain, their expression might be under complex regulatory control. It cannot be excluded, however, that some putative bacteriocin-encoding genes might actually function as inducing factors. Similar results were obtained for other genomes that were analyzed with BAGEL (see Supplementary Table S1).

Identification of new bacteriocins using ORF prediction

To investigate whether the omission of small ORFs in bacterial genomes results in the underestimation of the number of bacteriocin genes, the ORF prediction tool was applied on the genome sequence of *S.pneumoniae* TIGR4. The use of this new set of ORFs in BAGEL allowed the identification of one additional potential bacteriocin. It showed similarity to bacteriocin PlnB of *Lactobacillus plantarum*. Interestingly, this peptide was also identified in *S.pneumoniae* R6 after applying the same procedure (23), demonstrating the importance of re-annotation of an original genome sequence to be able to identify as many as possible putative bacteriocin genes.

Outlook

The power of BAGEL in annotating putative bacteriocin genes stems from the fact that it combines (i) all information on sequence motifs, characteristics and functions of the proteins involved in the biosynthesis of the putative bacteriocin, with (ii) the genetic context of the encoding genes and, (iii) our knowledge-based bacteriocin database. Owing to the enormous variety among the five main bacteriocin classes and their sub-classes, the predicted bacteriocins might still be false positives. Future insights in the interesting and rapidly developing field of bacteriocin research will help improving the BAGEL algorithm, as its database will be updated regularly with confirmed new bacteriocin information.

SUPPLEMENTARY DATA

Supplementary data are available at *NAR* Online.

ACKNOWLEDGEMENTS

REFERENCES

1. O'Sullivan,L., Ross,R.P. and Hill,C. (2002) Potential of bacteriocin-producing lactic acid bacteria for improvements in food safety and quality. *Biochimie*, **84**, 593–604.
2. Cotter,P.D., Hill,C. and Ross,R.P. (2005) Bacteriocins: developing innate immunity for food. *Nat. Rev. Microbiol.*, **3**, 777–788.
3. Brook,I. (2005) The role of bacterial interference in otitis, sinusitis and tonsillitis. *Otolaryngol. Head Neck Surg.*, **133**, 139–146.
4. Gillor,O., Nigro,L.M. and Riley,M.A. (2005) Genetically engineered bacteriocins and their potential as the next generation of antimicrobials. *Curr. Pharm. Des.*, **11**, 1067–1075.

5. McAuliffe,O., Ross,R.P. and Hill,C. (2001) Lantibiotics: structure, biosynthesis and mode of action. *FEMS Microbiol. Rev.*, **25**, 285–308.

6. Jack,R.W., Tagg,J.R. and Ray,B. (1995) Bacteriocins of Gram-positive bacteria. *Microbiol. Rev.*, **59**, 171–200.

7. Dirix,G., Monsieurs,P., Dombrecht,B., Daniels,R., Marchal,K., Vanderleyden,J. and Michiels,J. (2004) Peptide signal molecules and bacteriocins in Gram-negative bacteria: a genome-wide *in silico* screening for peptides containing a double-glycine leader sequence and their cognate transporters. *Peptides*, **25**, 1425–1440.

8. Eijsink,V.G., Axelsson,L., Diep,D.B., Havarstein,L.S., Holo,H. and Nes,I.F. (2002) Production of class II bacteriocins by lactic acid bacteria; an example of biological warfare and communication. *Antonie Van Leeuwenhoek*, **81**, 639–654.

9. Garneau,S., Martin,N.I. and Vederas,J.C. (2002) Two-peptide bacteriocins produced by lactic acid bacteria. *Biochimie*, **84**, 577–592.

10. Kemperman,R., Kuipers,A., Karsens,H., Nauta,A., Kuipers,O.P. and Kok,J. (2003) Identification and characterization of two novel clostridial bacteriocins, circularin A and closticin 574. *Appl. Environ. Microbiol.*, **69**, 1589–1597.

11. Riley,M.A. and Wertz,J.E. (2002) Bacteriocin diversity: ecological and evolutionary perspectives. *Biochimie*, **84**, 357–364.

12. Stein,T. (2005) *Bacillus subtilis* antibiotics: structures, syntheses and specific functions. *Mol. Microbiol.*, **56**, 845–857.

13. Pearson,W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.*, **24**, 307–331.

14. 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.

15. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

16. Guo,F.B. and Zhang,C.T. (2006) ZCURVE_V: a new self-training system for recognizing protein-coding genes in viral and phage genomes. *BMC Bioinformatics*, **7**, 9.

17. Besemer,J. and Borodovsky,M. (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.*, **33**, W451–W454.

18. Baerends,R.J., Smits,W.K., de,J.A., Hamoen,L.W., Kok,J. and Kuipers,O.P. (2004) Genome2D: a visualization tool for the rapid analysis of bacterial transcriptome data. *Genome Biol.*, **5**, R37.

19. Nes,I.F. and Johnsborg,O. (2004) Exploration of antimicrobial potential in LAB by genomics. *Curr. Opin. Biotechnol.*, **15**, 100–104.

20. Paik,S.H., Chakicherla,A. and Hansen,J.N. (1998) Identification and characterization of the structural and transporter genes for, and the chemical and biological properties of, sublancin 168, a novel lantibiotic produced by *Bacillus subtilis* 168. *J. Biol. Chem.*, **273**, 23134–23142.

21. Klaenhammer,T.R. (1993) Genetics of bacteriocins produced by lactic acid bacteria. *FEMS Microbiol. Rev.*, **12**, 39–85.

22. Tettelin,H., Nelson,K.E., Paulsen,I.T., Eisen,J.A., Read,T.D., Peterson,S., Heidelberg,J., DeBoy,R.T., Haft,D.H., Dodson,R.J. *et al.* (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*, **293**, 498–506.

23. Hoskins,J., Alborn,W.E.,Jr, Arnold,J., Blaszczak,L.C., Burgett,S., DeHoff,B.S., Estrem,S.T., Fritz,L., Fu,D.J., Fuller,W. *et al.* (2001) Genome of the bacterium *Streptococcus pneumoniae* strain R6. *J. Bacteriol.*, **183**, 5709–5717.

24. Marciset,O., Jeronimus-Stratingh,M.C., Mollet,B. and Poolman,B. (1997) Thermophilin 13, a nontypical antilisterial poration complex bacteriocin, that functions without a receptor. *J. Biol. Chem.*, **272**, 14277–14284.

25. Yamazaki,K., Suzuki,M., Kawai,Y., Inoue,N. and Montville,T.J. (2005) Purification and characterization of a novel class IIa bacteriocin, piscicocin CS526, from surimi-associated *Carnobacterium piscicola* CS526. *Appl. Environ. Microbiol.*, **71**, 554–557.

26. Twomey,D., Ross,R.P., Ryan,M., Meaney,B. and Hill,C. (2002) Lantibiotics produced by lactic acid bacteria: structure, function and applications. *Antonie Van Leeuwenhoek*, **82**, 165–185.

27. Nes,I.F. and Holo,H. (2000) Class II antimicrobial peptides from lactic acid bacteria. *Biopolymers*, **55**, 50–61.

28. Pons,A.M., Delalande,F., Duarte,M., Benoit,S., Lanneluc,I., Sable,S., Van,D.A. and Cottenceau,G. (2004) Genetic analysis and complete primary structure of microcin L. *Antimicrob. Agents Chemother.*, **48**, 505–513.