Published online 28 April 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W145–W148***
doi:10.1093/nar/gkn228

CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats

Ibtissem Grissa[1,*], Gilles Vergnaud[1,2] and Christine Pourcel[1]

[1]Univ. Paris-Sud 11, CNRS, UMR8621, Institut de Génétique et Microbiologie, 91405 Orsay and [2]DGA/D4S - Mission pour la Recherche et l'Innovation Scientifique, 7, rue des Mathurins, 00470 Armées, France

Received January 25, 2008; Revised April 6, 2008; Accepted April 11, 2008

ABSTRACT

Clustered regularly interspaced short palindromic repeat (CRISPR) elements are a particular family of tandem repeats present in prokaryotic genomes, in almost all archaea and in about half of bacteria, and which participate in a mechanism of acquired resistance against phages. They consist in a succession of direct repeats (DR) of 24–47 bp separated by similar sized unique sequences (spacers). In the large majority of cases, the direct repeats are highly conserved, while the number and nature of the spacers are often quite diverse, even among strains of a same species. Furthermore, the acquisition of new units (DR + spacer) was shown to happen almost exclusively on one side of the locus. Therefore, the CRISPR presents an interesting genetic marker for comparative and evolutionary analysis of closely related bacterial strains. CRISPRcompar is a web service created to assist biologists in the CRISPR typing process. Two tools facilitates the *in silico* investigation: CRISPRcomparison and CRISPRtionary. This website is freely accessible at http://crispr.u-psud.fr/CRISPRcompar/.

INTRODUCTION

The clustered regularly interspaced short palindromic repeat (CRISPR)-associated system (CASS) comprises the particular repeated element CRISPR itself, the promoter for its transcription (also called the leader) and a set of *cas* genes responsible for its maintenance and function (1,2). It is found in most Archea and 40% bacteria, and is linked to a mechanism of acquired resistance against bacteriophages (3). Some genomes harbour a significant number of CRISPRs [18 in *Methanocaldococcus jannaschii DSM 2661*with three different direct repeats (DRs)] (4). When different CRISPRs with the same DR are present in a genome, they have a very similar leader, generally different spacers, and only one is associated with *cas* genes (5). When CRISPRs from different CRISPR families exist in the same genome, one set of cas genes specific for each family is present. Finally, within a species, different strains may have different CRISPRs. The example of the three sequenced strains of *Streptococcus thermophilus* is very illustrative of this situation, since three CRISPRs were identified in this species but only strain LMD-9 possesses the three of them (4).

CRISPRs evolve either by deletion or acquisition of units (a DR and a spacer) following a mechanism proposed firstly by Pourcel *et al.* (6) and recently confirmed (7–9). In the majority of cases, new units are added at one end of the CRISPR adjacent to the leader, whereas motif deletions can occur randomly. The independent acquisition of the same spacer twice is possible but is not frequent and easily detected. Thus, the presence of identical spacers in the same CRISPR locus in distinct strains reflects shared ancestry.

The polymorphism of CRISPRs can be used for molecular typing. The standard and classical technology developed for *Mycobacterium tuberculosis* typing (10) is the spoligotyping, which consists in detecting the presence/absence of a range of spacers. This technique and other PCR-based typing methods have been applied in CRISPR genotyping to study other bacterial species (6,11–16).

We recently implemented a program (CRISPRFinder) allowing the identification of a CRISPR structure based on a thorough characterization of its components, i.e. the DR and the spacers (17). Using this program, public genome sequences are analysed and the extracted CRISPRs are stored into a database (CRISPRdb) (4). CRISPRFinder and CRISPRdb are accessible on the web together with different tools that assist in recovering spacers and DR sequences, and blasting them against Genbank.

We now report on the development of a new website dedicated to the comparison of CRISPRs between strains and the labelling of spacers when multiple alleles are analysed.

CRISPRcompar is freely accessible at http://crispr.u-psud.fr/CRISPRcompar/index.php.

*To whom correspondence should be addressed. Tel: +33 1 69 15 30 01; Fax: +33 1 69 15 66 78; Email: ibtissem.grissa@igmors.u-psud.fr

METHODS AND IMPLEMENTATION

CRISPRcompar is a friendly web resource offering tools to compare CRISPRs between strains of a given species or between closely related species, and to classify the spacers. Its core routines were developed in Perl under Debian Linux. It is composed of two main applications; CRISPRcomparison and CRISPRtionary. CRISPRcomparison identifies and compares the CRISPRs of two or more genomes (complete or partial sequences). It is particularly useful when strains of a species possess several CRISPRs for which positions on the genome might vary, as a result for instance of large-scale genome rearrangements, or of presence–absence polymorphism of CRISPR loci in the genomes of interest. The similarity criteria are based on having an identical consensus DR and similar flanking sequences. The flanking sequences are compared by the ClustalW alignment of the 200 bp adjacent sequences to the CRISPR with a threshold of 90% of similarity. In the majority of cases, when multiple CRISPRs with the same DR are present in a genome, only one flanking sequence is similar, the one corresponding to the leader.

CRISPRtionary lists the spacers from different alleles derived from the same CRISPR locus and annotates them in a polarized fashion. Such data will be produced for instance when investigating the diversity (evolution) of CRISPRs within a species by sequencing the locus in different isolates. This tool can then be used to automatically number spacers, produce a 'dictionary' or repertoire of spacers and code the alleles using this dictionary. CRISPRFinder is used to identify the DR and order the spacers according to the DR sequence. When sequencing PCR products, the first few nucleotides may be missed or the data may be of poor quality. In addition, the first, often partial and degenerated DR (up to 50% of differences have been observed) may be missed by CRISPRFinder in this context. For this reason, a filter exploring the existence of stretches of additional DR in the flanking sequence was added so as to correctly identify the first spacer. It consists in blasting the two halves of the DR against the remaining nucleotides of the allele sequence. Given the mechanism of acquisition of new spacers, we recommend to orientate the CRISPR such that the degenerated DR is located on the left extremity and the leader is on the right. These criteria are convenient to attribute increasing numbers to the spacers from left to right, according to their acquisition order, i.e. the more recently added spacer close to the leader will be given the highest number.

Input

The CRISPRcompar program automatically recovers from CRISPRdb all strains containing a CRISPR and proposes to compare each of them using the alphabetic list (alternatively, all strains from a given genus can be selected at once using the 'strain taxonomy browser'). To compare unpublished sequences and genomes, a private database on the model of CRISPRdb (4) must first be created (http://crispr.u-psud.fr/CRISPRcompar/private/). Additional sequences from the private database can then be added in the comparison. Once a selection of sequences has been performed, the 'compare' button leads to a page where it is possible to choose the strain that will be used as a reference for the CRISPRs annotation. At this step, it is also possible to remove or add sequences in the comparison. When several alleles of a given locus are present in the submitted sequences, their spacers can be annotated using CRISPRtionary. Fasta files containing sequenced CRISPR alleles can also be directly submitted to CRISPRtionary.

Output

For the CRISPRcomparison application, the result is shown in a table where CRISPRs are grouped. Figure 1 shows the result of the comparison of three *S. thermophilus* strains. Information is given on the CRISPR position and on the number of repeats (Figure 1A). A link to the corresponding CRISPR in CRISPRdb can be activated. When two or more alleles of a given CRISPR are found, the flanking sequences can be aligned and a link is provided to the second application 'CRISPRtionary' to annotate and classify the spacers. By activating the 'compare spacer' button a table is shown in which the CRISPR sequences are provided in fasta format (Figure 1B). At this step, it is possible to upload a previous dictionary of spacers to which the spacers of the new CRISPR alleles will be compared. If no pre-determined dictionary exists, one will be created in the following steps. With the FindCRISPR button, the CRISPRFinder program is used to identify DRs and spacers. Often more than one DR candidate will be proposed for several reasons. One is due to the existence of several possible DRs, especially with short sequences (less than four units) and another is due to the CRISPR orientation on the genome. Indeed, when the submitted alleles are in different orientations, two DR sequences will be proposed. Therefore, the user should select the appropriate consensus DR or introduce a DR sequence. The 'find spacer' button leads to a page where spacers are labelled (Figure 1C) and different files can be recovered: (i) different formats of text and tab-delineated text files representing the corresponding CRISPRs and spacers labels (AnnotFasta, AnnotFasta_CodedAlleles, Fasta_CodedAlleles, Table_Coded_Alleles), (ii) 'Spacers dictionary' which is a tab-delineated text file containing a catalogue of the found spacers and their labels and (iii) 'binary file': a tab-delineated text-file where columns represent the spacer labels and rows represent the queried alleles. For each CRISPR allele, a spacer will be given the '1' value when it exists and '0' when it is absent. The binary file is especially interesting for providing a spoligotyping-like profile of the CRISPR and to visually illustrate the spacer composition in the strains. The different files may be used in further studies such as the evolutionary analysis of the species according to the spacer organization in the different strains or for epidemiological purposes.

The last step may be added to improve the output; this is called the re-annotation step. It might be interesting when a collection of alleles has been analysed to re-annotate the spacers such that numbering is increasing

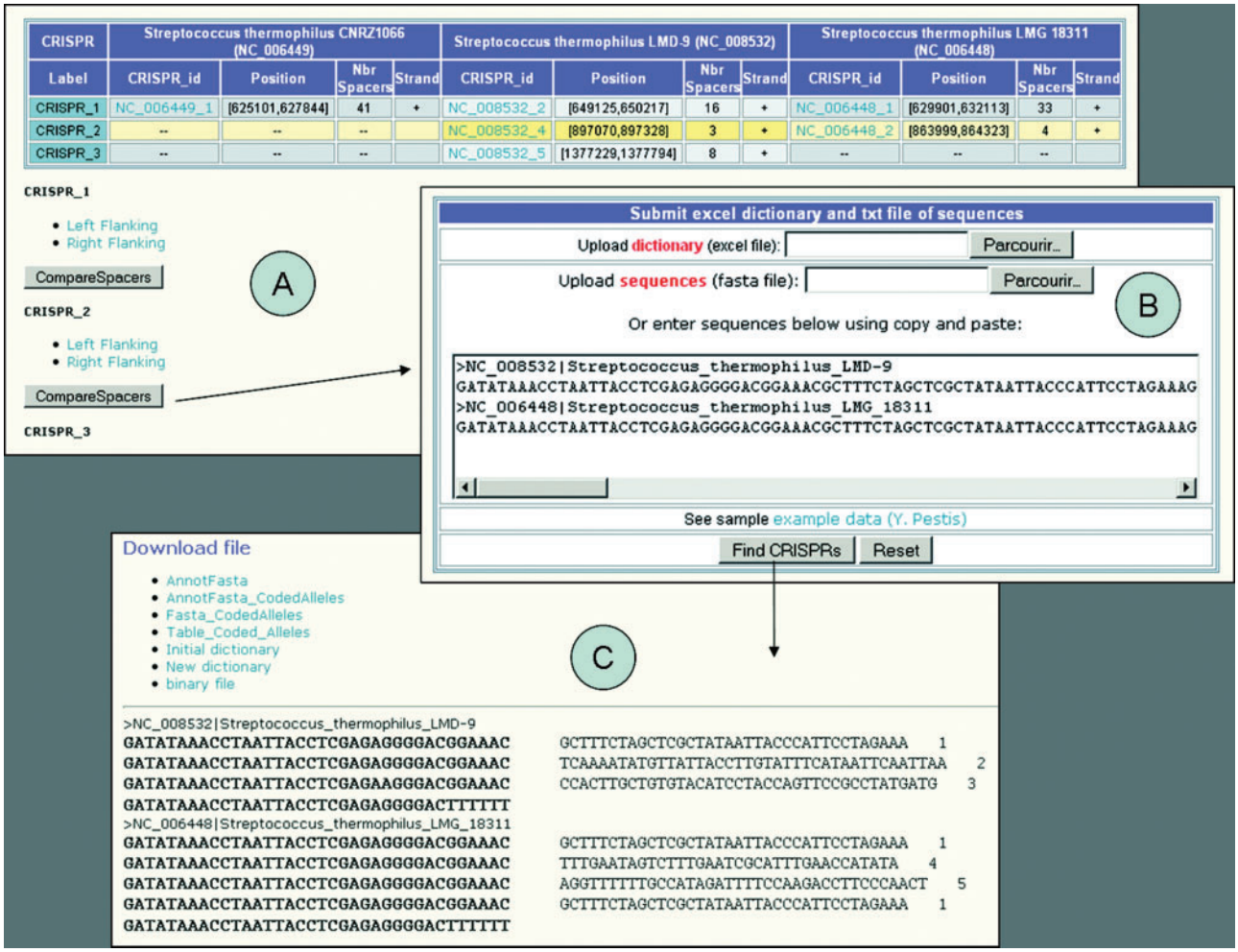

Figure 1. Example of CRISPRcompar and CRISPRtionary output using the three *S. thermophilus* genomic sequences (RefSeq: NC_006449, NC_008532, NC_006448). (**A**) Table showing the classification of the different CRISPRs. Three CRISPRs are identified, of which two are found in two or more strains. (**B**) CRISPR_2 sequences are submitted to the CRISPRtionary program. (**C**) The spacers are labelled and different files can be recovered.

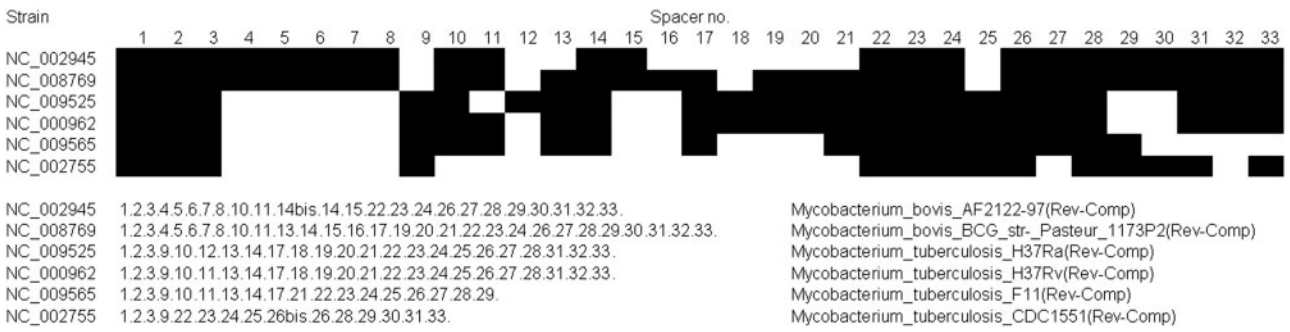

Figure 2. Schematic representation of the CRISPR repeats organization in four *M. tuberculosis* and two *M. bovis* strains. The binary file produced by CRISPRtionary, after the spacers have been re-annotated, is used to produce a figure in which the presence of a spacer is indicated by a dark square. The detail of the spacer composition for each strain is indicated in the bottom part of the figure.

starting from one end of the CRISPR. We propose that the oldest spacer, i.e. the one near the degenerated DR, when the later is identified, be given the label 1 and subsequent ones increasing numbers. The re-annotation tool modifies the labels such that all the labels inside an allele are in an increasing order and a new set of output files is produced. Sometimes, a duplication of one or several spacers may occur and in this case, the term 'bis' is added to the spacer label in the CRISPR code. On Figure 2 is shown the distribution and annotation of spacers in six members of the *M. tuberculosis* complex (MTBC). The binary file was converted into a diagram for

an easy comparison. The profile corresponds to the order of spacers described upon sequencing of a collection of alleles (18). However, and similarly to a real spoligotype, the presence in the same allele of two identical spacers is not indicated.

DISCUSSION AND CONCLUSIONS

The CRISPRcompar web server proposes a set of bioinformatic tools assisting biologists in the development and the setting up of a CRISPR genotyping scheme. In the pre-processing phase, the comparison of CRISPRs is mandatory and may be fulfilled using the CRISPRcomparison tool, which helps in selecting the most appropriate CRISPR loci and associated primers for the PCR amplification.

CRISPRcomparison allows the identification of families of strains that share a CRISPR, inside species with high genetic diversity or the identification of homologous CRISPRs within species containing multiple CRISPR loci. In the post-processing phase, the CRISPRtionary program is very interesting since it allows the user to easily compare multiple alleles of a CRISPR locus investigated in a collection of strains and to obtain pre-calculated files that may be directly used in clustering analysis. Many clustering methods are applicable and may provide a good clustering of the strains even if these methods usually do not take full advantage of the CRISPR rules of evolution, which could be used to better assess—in addition to forming groups of related strains—parental relations between taxa. The primary evolutionary events considered are motifs insertion and deletion. In the case of inactive (in terms of spacer acquisition) CRISPRs, only deletions are possible, and the Camin–Soakal (19) Parsimony model may be considered. In Camin–Soakal parsimony, two states are considered (0 and 1 for example), and no transition from derived state back to ancestral state is allowed. For an inactive CRISPR locus, the ancestral state is the presence of a unit and the derived state is unit absence; thus only deletion changes are allowed. Our future developments of CRISPRcompar will incorporate applications such as the MIX program of the package phylip (Felsenstein), which carries out the Camin–Soakal Parsimony method. It can be applied using the binary file with minor modifications.

ACKNOWLEDGEMENTS

The CNRS and Université Paris Sud 11 have funded this project. I.G. is supported by the TBChina EU project grant LSHPCT-2005-012166. Funding to pay the Open Access publication charges for this article was provided by Association Vaincre la Mucoviscidose.

Conflict of interest statement. None declared.

REFERENCES

1. Makarova,K.S., Grishin,N.V., Shabalina,S.A., Wolf,Y.I. and Koonin,E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct*, **1**, 7.
2. Sorek,R., Kunin,V. and Hugenholtz,P. (2008) CRISPR – a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.*, **6**, 181–186.
3. Deveau,H., Barrangou,R., Garneau,J.E., Labonte,J., Fremaux,C., Boyaval,P., Romero,D.A., Horvath,P. and Moineau,S. (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.*, **190**, 1390–1400.
4. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinform.*, **8**, 172.
5. Jansen,R., Embden,J.D., Gaastra,W. and Schouls,L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, **43**, 1565–1575.
6. Pourcel,C., Salvignol,G. and Vergnaud,G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, **151**, 653–663.
7. Lillestol,R.K., Redder,P., Garrett,R.A. and Brugger,K. (2006) A putative viral defence mechanism in archaeal cells. *Archaea*, **2**, 59–72.
8. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.
9. Tyson,G.W. and Banfield,J.F. (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.*, **10**, 200–207.
10. Groenen,P.M., Bunschoten,A.E., van Soolingen,D. and van Embden,J.D. (1993) Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol. Microbiol.*, **10**, 1057–1065.
11. Mokrousov,I., Limeschenko,E., Vyazovaya,A. and Narvskaya,O. (2007) *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci. *Biotechnol. J.*, **2**, 901–906.
12. Mokrousov,I., Narvskaya,O., Limeschenko,E. and Vyazovaya,A. (2005) Efficient discrimination within a *Corynebacterium diphtheriae* epidemic clonal group by a novel macroarray-based method. *J. Clin. Microbiol.*, **43**, 1662–1668.
13. Hoe,N., Nakashima,K., Grigsby,D., Pan,X., Dou,S.J., Naidich,S., Garcia,M., Kahn,E., Bergmire-Sweat,D. and Musser,J.M. (1999) Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. *Emerg. Infect. Dis.*, **5**, 254–263.
14. Schouls,L.M., Reulen,S., Duim,B., Wagenaar,J.A., Willems,R.J., Dingle,K.E., Colles,F.M. and Van Embden,J.D. (2003) Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *J. Clin. Microbiol.*, **41**, 15–26.
15. DeBoy,R.T., Mongodin,E.F., Emerson,J.B. and Nelson,K.E. (2006) Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences. *J. Bacteriol.*, **188**, 2364–2374.
16. Vergnaud,G., Li,Y., Gorge,O., Cui,Y., Song,Y., Zhou,D., Grissa,I., Dentovskaya,S.V., Platonov,M.E., Rakin,A. *et al.* (2007) Analysis of the three *Yersinia pestis* CRISPR loci provides new tools for phylogenetic studies and possibly for the investigation of ancient DNA. *Adv. Exp. Med. Biol.*, **603**, 327–338.
17. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.
18. van Embden,J.D., van Gorkom,T., Kremer,K., Jansen,R., van Der Zeijst,B.A. and Schouls,L.M. (2000) Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J. Bacteriol.*, **182**, 2393–2401.
19. Camin,J. and Soakal,R. (1965) A method for deducing branching sequences in phylogeny. *Evolution*, **19**, 311–326.