Chapter 2

How to Identify CRISPRs in Sequencing Data

Christine Drevet and Christine Pourcel

Abstract

Clustered regularly interspaced short palindromic repeats (CRISPRs) are DNA sequences composed of a succession of repeats (23–50 bp long) separated by unique sequences called spacers. CRISPRs together with a set of genes called *cas* for CRISPR associated, constitute a defence mechanism against invasion by foreign sequences. We describe protocols and bioinformatics tools that allow the identification of CRISPRs, their comparison and their component determination (the direct repeats and the spacers). A schematic representation of the spacer organization can be produced, allowing an easy comparison between strains.

Key words: CRISPR, Genotyping, Bacteriophage, Database, Spacer, Phylogeny

1. Introduction

Clustered regularly interspaced short palindromic repeats (CRISPRs) loci typically consist of the succession of 23–50 bp direct repeats (DR), separated by variable and non-repetitive sequences called spacers (see Fig. 1). A CRISPR generally possesses at one end a degenerated DR (DR′) and at the other end a complete DR immediately followed by a sequence called the leader, which acts as a promoter (see Fig. 1a). In a single genome several CRISPRs with the same DR can be found, but only one is associated with a group of genes called *cas* (for CRISPR-associated) (see Fig. 1b) (1, 2). The CRISPR/cas system has been identified in a broad range of prokaryotic species, 85% of Archaea and 48% of Bacteria (3).

In the majority of cases, the spacers, when identified, happen to be fragments of bacteriophages or plasmids (4, 5). Different observations suggest that the CRISPR/cas system constitutes a defence system against foreign sequences. The currently sequenced

Kenneth C. Keiler (ed.), *Bacterial Regulatory RNA: Methods and Protocols*, Methods in Molecular Biology, vol. 905,
DOI 10.1007/978-1-61779-949-5_2,

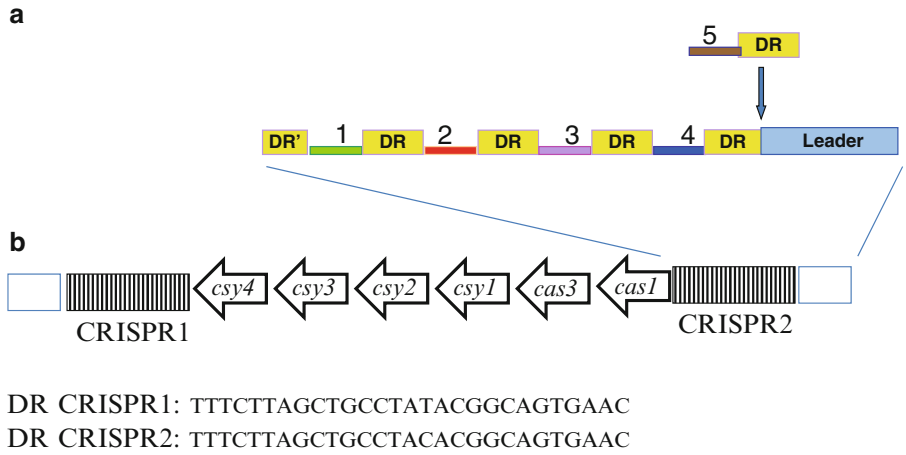

Fig. 1. Schematic structure of a CRISPR and the *Yersinia* subtype CRISPRs in *Pseudomonas aeruginosa*. (**a**) DRs are shown as *yellow boxes* and DR′ represent the degenerate DR on one side of the CRISPR. On the other side is the sequence called leader, which acts as a promoter. Spacers are shown with different colors and are numbered according to the order of their acquisition. A new spacer is added next to the leader after duplication of the last DR element. (**b**) Schematic organization of the CRISPR/cas locus. A set of six *cas* genes are framed by two CRISPR structures. The DR sequences of CRISPR1 and CRISPR2 differ at one nucleotide.

CRISPR structures listed in CRISPRdb (3) show an important variability in the nature of DRs and the number of motifs between and inside species. The largest observed CRISPR to date possesses 588 spacers in *Haliangium ochraceum* DSM 14365. Within a species, strains may or may not possess a given CRISPR (with a particular DR) and the number and nature of spacers may vary considerably. This diversity suggests that the CRISPR structure is continuously evolving, either through the addition of new motifs (a DR and a spacer) or by interstitial deletion of one or several motifs through recombination between two DRs. New motifs are added to the CRISPR in a polarized manner by duplicating the DR next to the leader and adding a new fragment of DNA (4, 6) (see Fig. 1a).

Correctly identifying CRISPRs in a bacterial or archaeal genome is not as straightforward as might be expected. A major challenge is that other types of repeats could be misidentified as CRISPR, so that it is necessary to identify features distinguishing CRISPRs from other repeated sequences. In addition, rules for CRISPR identification deduced from known CRISPRs may be too restrictive to find CRISPR sequences in new datasets. It is difficult to define parameters that will encompass the characteristics of all CRISPRs (including the smaller ones) and, in particular, faithfully define their DRs boundaries. Several programs have been developed to specifically search for CRISPRs, including CRISPRfinder (7), which is used to build the database CRISPRdb (3). Both CRISPRfinder and CRISPRdb are available at the CRISPR web

server site http://crispr.u-psud.fr/crispr/. Recent improvements were made to the CRISPR web server by giving access to information on *cas* genes present in the analyzed genomes, and by simplifying the comparison of CRISPR alleles. The database, which is being regularly updated and manually curated, provides files that help in analyzing the diversity of CRISPR elements. The methods below explain how to use these tools to identify CRISPRs in new sequence data, analyze CRISPR components, and compare CRISPRs in different genomes.

2. Materials

The web-based tools necessary for CRISPRs identification and comparison are freely accessible at http://crispr.u-psud.fr.

2.1. CRISPRs Database

The CRISPR web server hosts the CRISPRdb (a MySQL open source database), storing the CRISPRs content from reference genomes of 118 Archaea and 1,582 Bacteria as of April 2012. The database is regularly updated and curated by the authors.

2.2. User Data Analysis

2.2.1. CRISPRfinder

CRISPRfinder is used to identify and characterize CRISPR-like structures. Nucleic acid sequences in FASTA format can be pasted into input text areas or uploaded from a file on the local machine. Multi-sequence files are also allowed by the CRISPRfinder program and each sequence will be treated independently. The output is primarily web-based and viewable on the website. In addition, results can be downloaded as text files.

2.2.2. CRISPRcompar

CRISPRcompar is used to identify allelic CRISPRs from different strains according to their genomic position.

Strains to be compared are selected in the public and private database. Output is viewable on the website. It can be imported into CRISPRtionary for further analysis of spacers.

2.2.3. CRISPRtionary

CRISPRtionary is used to compare spacer arrangement of CRISPRs. The data are a set of allelic sequences of a given CRISPR locus. The output is viewable on the website. In addition, results can be downloaded as text files or as tabulated text files that can be opened with any table viewers, including Excel.

2.3. User Data Storage

Users can create a private account to store (in a MySQL database) and further analyze their own data.

3. Methods

The different tools available on the CRISPR web server allow three main applications: CRISPRs can be identified in users' data directly by submitting sequences to CRISPRfinder (1), CRISPRs can be identified in reference genomes by browsing the CRISPRdb database (2), and CRISPRs in different strains can be compared using the CRISPRcompar and CRISPRtionary tools (3).

3.1. Finding CRISPRs

The smallest CRISPRs detected by CRISPRfinder consist of two DRs (a complete and a degenerated one) separated by a spacer (see Note 1). Large CRISPRs can contain several hundred repeats. The presence of a CRISPR in a strain does not imply its existence in all the members of the species.

1. Submit nucleic acid sequences to the CRISPRfinder program online (http://crispr.u-psud.fr/Server/) by uploading a sequence file or pasting the sequences into the dialog box.
2. Click the FindCRISPR button to launch the CRISPRfinder program with the default parameters (see Note 2). CRISPRfinder advanced version, available from the main CRISPRfinder form, allows modifying the parameters to refine the search (see available options in Note 2). The resulting pages list the identified CRISPRs for each submitted sequence, and separate them into "confirmed" and "questionable" (see Fig. 2; Note 3). Each

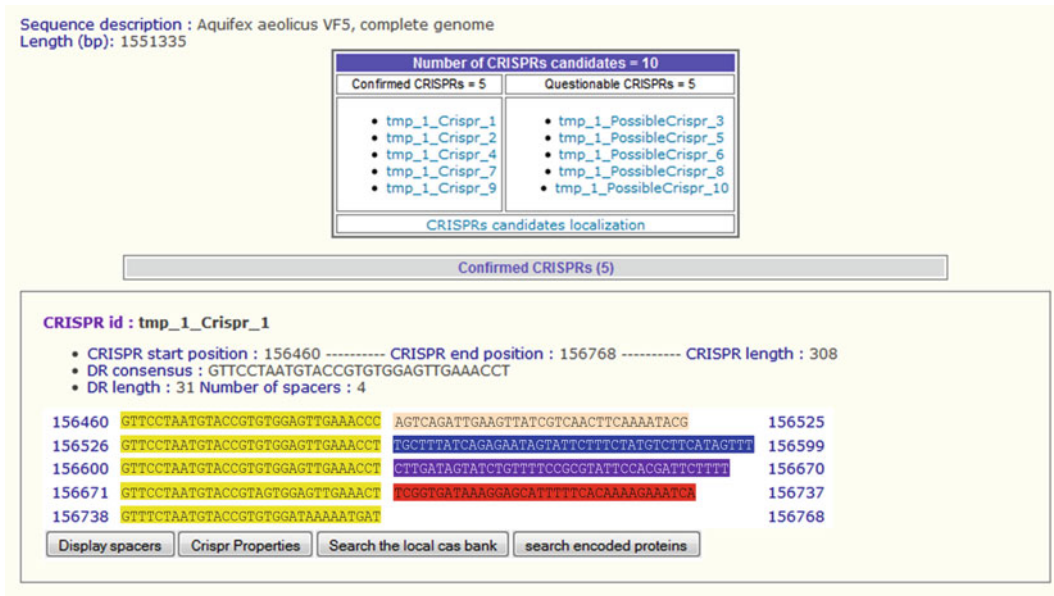

Fig. 2. Output of the CRISPRfinder program. Confirmed and questionable CRISPRs are listed and each of them is depicted in a schematic representation where the DRs are shown in *yellow* and the spacers with different colors.

identified structure is depicted where the DR is shown in yellow and the spacers are shown with different colors.

3. By clicking on the different buttons under the identified CRISPR, one can display the FASTA formatted spacer list (Display spacers), and the CRISPR properties file (CRISPR properties) which are the output of the CRISPRfinder stand-alone program. The CRISPR properties include DR size, DR consensus sequence, sequences and coordinates of the spacers.
4. Identification of *cas* genes near a questionable CRISPR can provide evidence that the CRISPR is legitimate, as each organisms that contains bona fide CRISPRs should contain a set of *cas* genes. To search for *cas* genes near an identified CRISPR, click the "Search the local *cas* bank" button. Similarities between the submitted data and *cas* genes from a large number of species in the vicinity of the CRISPR (±10,000 bp) will be identified (see Note 4).
5. The "search encoded proteins" function checks whether the CRISPR sequence corresponds to an open reading frame. This tool may help in eliminating structures that are not true CRISPRs, because there is no documented case of CRISPRs including functional proteins (see Note 5).

3.2. Analysis of CRISPR Components

Several features to analyze the CRISPRs DR and spacers are gathered in the pink box on the top right of the page.

1. To identify putative protospacers for CRISPRs, click the "Search protospacers using BLAST" button. The spacers of each CRISPR will be displayed, and BLASTN searches can be performed with each spacer to find similar sequences in the Genbank nr database (see Note 6).
2. Sequences surrounding the CRISPR on each side can be recovered using the "Extract the flanking sequences" button. It is then possible to launch BLASTN searches with these flanking sequences in the Genbank nr database. This feature is helpful when checking whether a CRISPR region is present in other genomes.
3. Clicking on "Search CRISPRs with identical DRs" leads to a table listing strains and CRISPRs identity numbers (Ids) (see Note 7) found in the CRISPR database. These CRISPRs can be visualized by clicking on each Ids. Questionable CRISPRs from Subheading 3.1, step 3, bearing numerous spacers can be validated as true CRISPRs when shared spacers are identified.
4. Sometimes there are small natural variations in DR sequences, even between closely related species. The BLAST CRISPRdb link in the left main menu should be used to identify related DRs from the database.
5. Finally text files of spacers and CRISPR properties can be downloaded.

3.3. CRISPRs Public Database

1. The presence of a CRISPR in a public sequenced genome can be assessed by consulting the CRISPRdb database (http://crispr.u-psud.fr/crispr/). Strains marked in pink color possess confirmed CRISPR, whereas those in gray have only questionable structures and those in yellow have no CRISPR (see Note 8). By default, the strains are presented alphabetically, but choosing the "View the strains taxonomy browser" link shows the strains according to taxonomy (see Note 9).
2. Clicking on a strain leads to a table listing the sequences present in its genome (chromosome and plasmid) and the associated CRISPRs (see Fig. 3a).
3. At this stage it is possible to view annotated *cas* genes in the genomes which harbor a CRISPR (see Note 10). Because the *cas* genes are typically not systematically annotated in genome sequences, a complementary search in the local database of CAS proteins is recommended. Gene members of four CAS protein families, designated Cas1 to Cas4, are expected in the

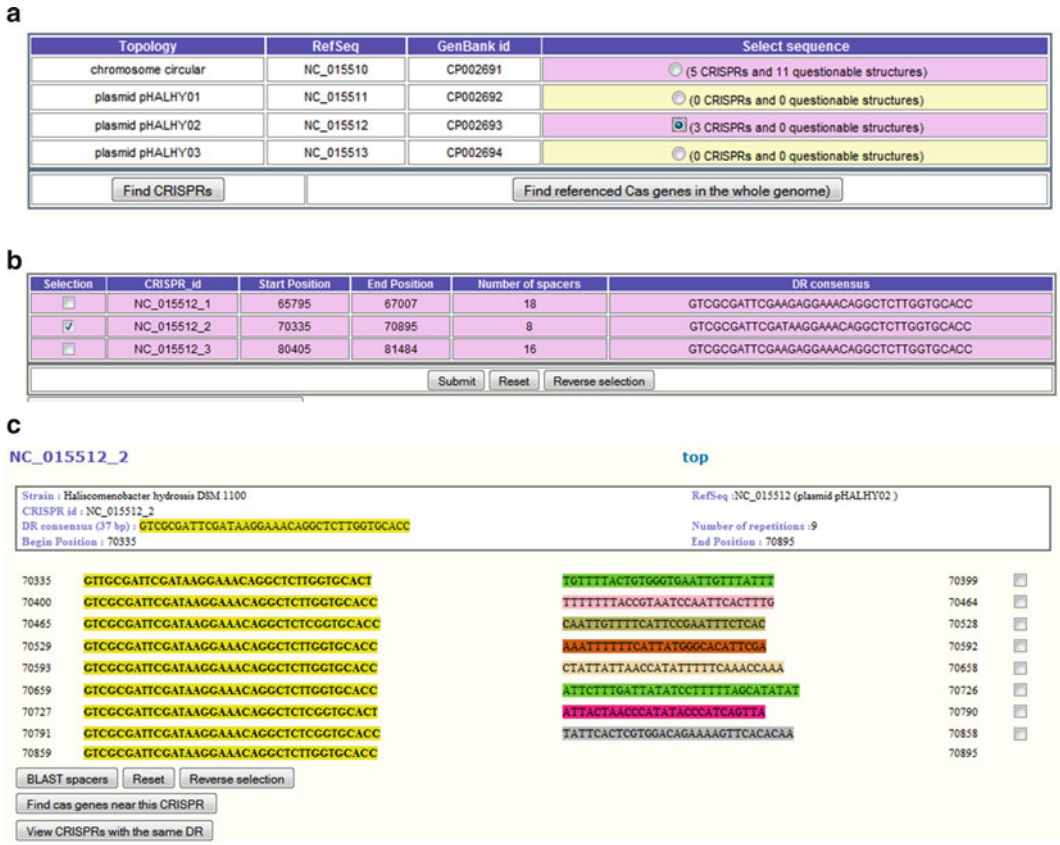

Fig. 3. Output of a CRISPRdb query. (**a**) Each sequence contained in the queried strain is listed with its RefSeq number, GenBank id and the number of CRISPRs found. (**b**) The list of CRISPRs found in the selected sequence is shown, with its position and the DR consensus sequence. (**c**) The selected CRISPR is shown.

vicinity of at least one structure in strains harboring true CRISPRs (see Note 3).

4. Selected CRISPRs can be viewed by clicking on the submit button (see Fig. 3b). The schematic representation is similar to that of CRISPRfinder and tools are available to BLAST spacers against Genbank, search for clustered *cas* genes and view CRISPRs with the same DR (see Fig. 3c).
5. Direct Repeats and spacers lists of "confirmed" CRISPRs are built from the CRISPR database after each update. They are viewable and downloadable in "CRISPR utilities" and available for BLAST searches at the "BLAST CRISPRS" menu item.

3.4. My CRISPRdb

1. It is possible to run CRISPRfinder on submitted sequences and to store the results in a dedicated database hosted by the CRISPR server. To create a private account, an email address and a password must be provided.
2. Sequences can be submitted to CRISPRfinder using the same form as in the main CRISPRfinder page.
3. The identified CRISPR structures can be saved after providing a sample name.
4. To access the database click on the "Consult your private database button." The private data browser is similar to that of the main CRISPRdb. It is also possible to compare CRISPRs in both databases using the CRISPRcompar tool (see Subheading 3.5).

3.5. CRISPRs Comparison

CRISPR polymorphism is generally observed within a species and this feature may be used to compare strains and to provide phylogenetic information. The following tools help in identifying CRISPR loci and classifying spacers.

1. When the genome sequences of several strains are available for a given species, and when each strain possesses several CRISPRs, it is important to be able to classify the different CRISPRs, particularly as their position on the genome might vary. The CRISPRcompar tool is based on the presence of identical DRs and similar flanking sequences (100–70% mismatch threshold is accepted) (see Note 11). This function is particularly useful when several CRISPRs are present in a single genome.

   Three optional forms are available when connected to the CRISPRcompar page. The "View all the genomes harbouring CRISPRs" form lists all the sequences of the CRISPRdb that contain true CRISPRs in alphabetical order. Another way to compare CRISPR loci is to activate "View the strains taxonomy browser" which recovers from CRISPRdb all members of a genus containing a CRISPR and allows comparison of each

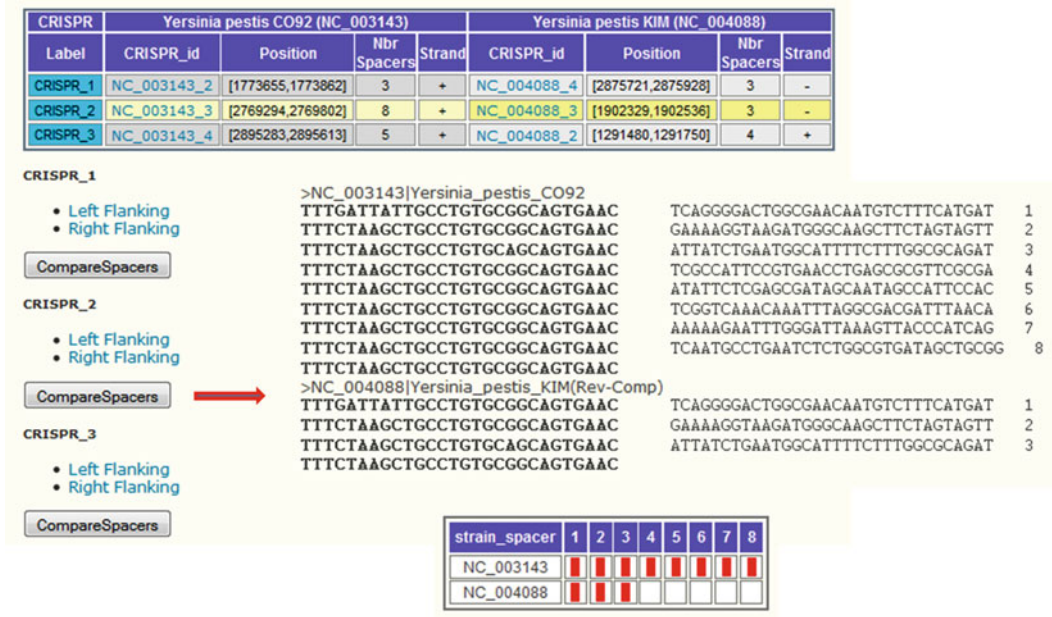

Fig. 4. Output of a CRISPRcompar query. The different CRISPRs of a given locus are compared by giving access to their flanking sequences and to their spacer's list.

of them. Additional strains can be added to the comparison. Finally logging into a private database leads to a page where it is possible to activate the link "CRISPRs comparison". Thereafter, the list of genomes present in the private database will be available together with the CRISPRdb list for comparing CRISPRs.

2. Select genomes, then activate the compare button. The resulting page shows the selected genomes, which can be retained or deleted from the comparison. Additional genomes can be added at that stage.
3. Select one of the genomes to serve as a reference for labeling the different CRISPR loci, and click the compare button. The resulting page shows the output of the comparison in a table in which CRISPR loci are listed under each sequence with their coordinates and number of spacers (see Fig. 4).
4. When two sequences possess a common CRISPR locus, their spacers can be compared by clicking on the "Compare Spacers" button. Activating this button will lead to the CRISPRtionary program in which spacers of submitted CRISPR sequences can be compared (see Subheading 3.6).
5. After getting the list of CRISPR loci in compared genomes, it is possible to view an alignment of sequences flanking the CRISPRs of a given locus by clicking on the "Left flanking" or

"Right flanking" links (see Note 11). Alignments between 5′ and 3′ sequences of CRISPRs are also directly available with the FlankAlign tool. In that case the user can choose the length of the regions to align (CRISPRdb sequences).

3.6. CRISPRtionary

1. After activating this page, upload or paste into the frame a list of sequences in FASTA format in order to create a repertoire of spacers, annotate them and order them (see Fig. 4). The spacers dictionary is an excel-formatted file with a specific format (examples are provided below the submission form, see Note 12). A dictionary containing the spacers identified in the submitted sequences will be created by the program if you do not specify any at this stage.
2. After activating the "Find CRISPRs" button click on continue if no dictionary was uploaded and the next page will show the CRISPRs identified in each of the submitted sequences.
3. To number the spacers it is necessary to select a DR sequence either from the "List of candidate DRs" or by introducing a sequence (see Note 13).
4. The "Find spacers" button leads to a page where each spacer has been given a number and different tables are available. Known spacers are numbered with the keys of the uploaded dictionary of spacers. New spacers will be given a new Id number and an updated dictionary will be created. The data is provided in the form of a tabulated text file, which can be stored and used for further analysis of new alleles.
5. Choose "see the html version" to see dictionaries as tables of spacers and strains.
6. The "Re-annotate Spacers" tool allows easy assessment of tentative phylogenetic relationships between alleles. The graphical representation of ordered spacers at the top of the page shows the spacer organization.

4. Notes

1. These structures have too low complexity to be identified de novo in sequences. However they are detected by comparison using the CRISPRtionary tool (as in *Yersinia pestis* Nepal 516, Accession number: NC_008149, coordinates: 233155-2331643).
2. Default parameters (as described in Grissa et al. (7) and in the web server "page manual" link) used in the regular CRISPRfinder are the following: In a first step, a search for two consecutive DRs with a length of 23–55 bp and an internal

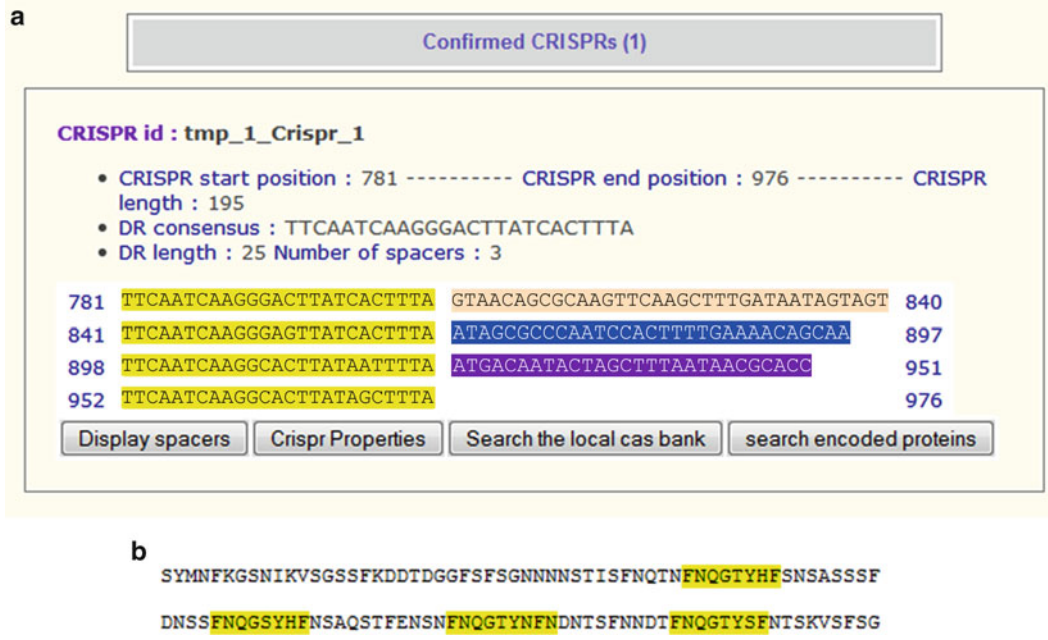

Fig. 5. CRISPR-like sequence present in the Helpy-906 gene coding for cytotoxin VacA in *Helicobacter pylori*. (**a**) CRISPRfinder output using the all gene sequence. (**b**) Translation of the CRISPR-like region.

spacer size of 25–60 bp is performed. At this stage, one nucleotide mismatch is allowed between repeats. In the advanced version, these parameters are modifiable in the "Maximal repeat" section of the form. In a second step, the recovered sequences are clustered to build the final CRISPRs. At this stage, the spacer size should be from 0.6- to 2.5-fold the DR size. The default values of the allowed mismatch between DRs are 20% for the internal DRs and 33.3% for the degenerated one. The similarity between spacers should not exceed 60% in order to discard tandem repeats. These values are modifiable by the user in the "CRISPR properties" section of the advanced form.

3. Small CRISPRs harboring two or three DRs are classified as "Questionable." In such structures there are not enough DR copies to accurately define a consensus. Longer CRISPRs are annotated as "Confirmed." However a critical examination of the CRISPRfinder output is necessary because some structures may be misidentified as CRISPR if their characteristics fit all the CRISPRfinder program specifications. For example, portions of some genes consist of well-conserved sequences separated by more diverged sequences (see Fig. 5). A very low rate of mismatches between the internal DRs of a CRISPR is observed for most of the organisms analysed in the CRISPRdb.

4. A local database has been implemented that contains amino acid sequences of Cas proteins present in the Uniprot database (http://www.uniprot.org/). BLASTX searches against this database are made in order to find potential *cas* genes locations in DNA sequences. The presence of *cas*-like genes in cis highly validates CRISPRs-like structures but *cas* genes also act in trans in many cases.
5. CRISPR structures are not coding regions although they are made of gene fragments (the spacers) and repeats that may happen to produce an open reading frame. A putative CRISPR might in fact turn out to be part of a protein-coding sequence with repeated elements.
6. The sequence from which a spacer originated is called the protospacer. It is usually localized on a viral or plasmid genome although some spacers have been shown to originate from bacterial genes.
7. The CRISPR Id is made of the sequence NC number followed by a number corresponding to the CRISPR rank identified by the program in that sequence. Some CRISPR numbers may be missing in the database if they have been deleted by filters or during the curation process.
8. CRISPRdb stores the characteristics and location of CRISPRs identified de novo by the CRISPRfinder program in all the Bacteria and Archaea sequences (chromosome and associated plasmids) from the RefSeq database released at the ftp site of the NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Several additional steps allow comparing candidates to CRISPRs already present in the database in order to convert some questionable structures into confirmed ones and to discard others. CRISPRdb is periodically updated and on this occasion each new CRISPR is critically observed. An administrator interface is dedicated to the sorting and to manual correction before final validation of the new version of the production database. A CRISPR harboring highly variable DRs is specifically examined and checked for the presence of long open reading frames. If a potential CRISPR corresponds to an ORF in Genbank, the CRISPR is deleted from the database. Indeed, some coding sequences possess exact repeats separated by sequences with a high degree of divergence. Similarly, the CRISPR is deleted when the spacer length is highly heterogenous, unless a set of *cas* genes is present in the vicinity.
9. Taxonomic information are those of PublicHouse, a publicly queryable relational databases (see http://biowarehouse.ai.sri.com/PublicHouseOverview.html for information). Genomes with missing information are not listed in the Taxonomy browser.

10. In the current version, 85% of archaea and 48% of bacteria possess one or several CRISPRs, and in 70% of all genomes possessing a CRISPR there is at least one *cas* gene referenced in the sequence annotations (chromosome or associated plasmids). When several CRISPRs are present in a single genome, a set of *cas* genes is generally clustered with at least one CRISPR. Up to three different CRISPR/Cas systems were observed in some genomes, such as in *Thermincola* sp.
11. The comparison program is performing alignments of the 200 bp regions flanking each CRISPR. The alignment is produced by the clustalW software (ktuple = 2, matrix = BLOSUM).
12. The dictionary must have three columns to be read by the program. The first column must contain numeric data and corresponds to the initial spacer keys that will appear in the "Spacers annotation" page at the right of each spacer. The second column, AnnotatedSpacer, must contain at least a string separated from a number by ":" (e.g., Sthermophilus:1). When the program is run, information about the strains that contain a given spacer and the position of that spacer in those strains are added in that field. The format of the resulting field will look like <strainA>:<spacer position in strainA>_<strainB>:spacer position in strain B>_<strainC> and so on. The third column contains the spacer sequences. When the re-annotate spacer program is launched, the spacer keys are changed to fit with the order of the spacers in the different strains CRISPRs.
13. The candidate DRs determined in each submitted sequence might differ at the last nucleotide, in particular when CRISPRs with a small number of motifs (a DR and a spacer) are analyzed. Therefore it is necessary to select a consensus DR, which is present in all the analyzed sequences. CRISPRtionary is particularly interesting to identify CRISPRs with a single spacer when one of the DR is degenerated. For example, the *Y. pestis* Nepal 516 strain possesses a CRISPR with one perfect DR and a truncated one separated by a unique spacer. That particular locus is allelic to a multi-DR locus in other *Y. pestis* strains. By imposing the use of the DR in CRISPRtionary it is possible to view the single spacer CRISPR.

References

1. Haft DH, Selengut J, Mongodin EF, Nelson KE (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol 1:e60
2. Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science 327:167–170
3. Grissa I, Vergnaud G, Pourcel C (2007) The CRISPRdb database and tools to display

CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8:172

4. Pourcel C, Salvignol G, 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
5. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol 60:174–182
6. Lillestol RK, Redder P, Garrett RA, Brugger K (2006) A putative viral defence mechanism in archaeal cells. Archaea 2:59–72
7. Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35:W52–W57

http://www.springer.com/978-1-61779-948-8

Bacterial Regulatory RNA
Methods and Protocols
Keiler, K.C. (Ed.)
2012, XI, 333 p., Hardcover
ISBN: 978-1-61779-948-8
A product of Humana Press