BIOINFORMATICS

Vol. 00 no. 00 2009
Pages 1–2

CRISPI : a CRISPR Interactive database

Christine Rousseau [1,*], Mathieu Gonnet [2], Marc Le Romancer[2], Jacques Nicolas [1,*]

[1]IRISA-INRIA, Campus de Beaulieu, 35042 Rennes cedex, France
[2]UMR 6197 Microbiologie des environnements extrêmes, technopôle Brest-Iroise, BP 70 29280 Plouzané, France.

Preprint HAL

ABSTRACT

Summary: The CRISPR genomic structures (Clustered Regularly Interspaced Short Palindromic Repeats) form a family of repeats that is largely present in archea and frequent in bacteria. Starting from a formal model of CRISPR using very few parameters, a systematic study of all their occurrences has been achieved in all available genomes of Archaea and Bacteria. It results in a relational database, CRISPI, including a complete repertory of associated *CAS* genes. A user-friendly web interface with many graphical tools and facilities allows extracting results, finding out CRISPR in personal sequences or calculating sequence similarity with spacers.

Availability: CRISPI free access at http://crispi.genouest.org

Contact: Jacques.Nicolas@inria.fr

1 INTRODUCTION

A remarkable regular structure made up of a skeleton of repeats holding a set of highly variable short sequences has been recognized several times in prokaryotic genomes under different names in the literature (TREP, SPIDR, SRSR...) and is called CRISPR since 2002 (for Clustered Regularly Interspaced Short Palindromic Repeats) (Barrangou *et al.*, 2007; Sorek *et al.*, 2008). The structure contains generally 4 to 10 direct repeats ranging in size from 25 to 45, separated by spacers of similar length containing specific genomic material that is not present elsewhere in the genome and has been likely imported from plasmids or viruses. CRISPR are present in all but 6 archaeal species and half of bacteria. Since they are expected to play an important role in prokaryotic adaptative immunity and may serve as specific markers, it is highly desirable to have dedicated identification tools and regularly updated databases available. Several computational methods have been developed to predict CRISPR using a more or less explicit model introducing many parameters filtering the allowed number of elements, sizes and distances between elements of the structure, mismatches between units (Bland *et al.*, 2007; Edgar., 2007; Grissa *et al.*, 2007)... The best source of data on CRISPR has been designed in 2007 by I. Grissa, G.Vergnaud and C. Pourcel (Grissa *et al.*, 2007) with the most recent release in december 2008. We have tried to improve this setting with a simpler CRISPR model and several new utilities.

2 IMPLEMENTATION

2.1 Identification of CRISPR

The usual specification of CRISPR, based on limited empirical data instead of biological functional constraints, remains too informal to be helpful in systematic studies: *CRISPR are repeated structures composed of exact repeat sequences 24 to 48 bases long separated by unique spacers of similar length*(Kunin *et al.*, 2007).

In fact most CRISPR include altered repeats and spacers are occasionally repeated inside a same structure and sometimes even in different CRISPR inside a same chromosome. Some authors give more details on the structure: repeats are supposed to exhibit a kind of dyadic symmetry but as more data are available this characterization becomes questionnable; A leader sequence is often mentionned before the train of repeats, but it is only defined as an A/T rich region and seems to lack for some CRISPR. Since the existence of a skeleton seems the only tangible fact for CRISPR and since we try to minimize a priori assumptions, we have chosen to only base the search on the existence of *a periodic spaced suite of units (at least four units) that is not a tandem repeat. Maximal repeats* have largely been used for the detection of relevant repeats and applied on the search for units (Grissa *et al.*, 2007). But short words such as those that appear in CRISPR can occur at a frequency comparable to random words of similar size. We have introduced locality restrictions on the notion of maximal repeat reflecting the kind of repeats that are found in CRISPR: first, each cluster of occurrences has a bounded size; second, only maximal repeats with at least one occurrence that is not covered by a larger repeat are kept. We produce putative units by clustering such overlapping local maximal repeats. Actually, we do not fix any value for the size of units or spacers, and we do not require units to be identical inside a given CRISPR (the minimal identity percentage to the consensus is however fixed to 60% in order to avoid spurious structures).

All bacterial and archaeal genomes have been downloaded from the NCBI FTP Server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The CRISPR have been searched for using the method sketched below implemented in C and Java (v. 1.5.0_12) and results have been stored into a MySQL (v. 4.1.12) database. All web pages are implemented using PHP (v. 4.3.9).

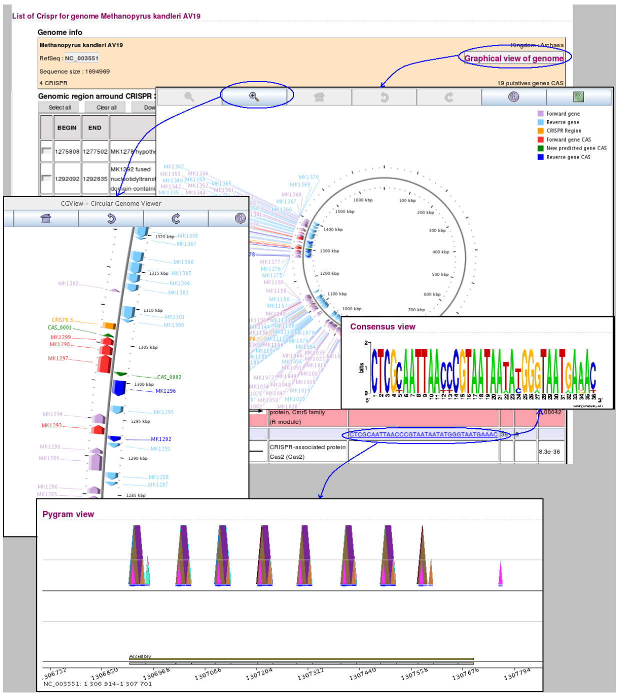
2.2 Access the CRISPI database

The main page of CRISPI offers three search forms: consult the content of the database, Blast a personal sequence against the database and find CRISPR structures in a personal sequence.

2.2.1 Consult the database CRISPI allows viewing all CRISPR found in Archaea and Bacteria genomes. Microbial genomes can

[*] to whom correspondence should be addressed

be easily selected by accession number, by entering the genome name (or a part of it) or by selecting a genome in the genome list (alphabetical order) or in the taxonomy browser. Once the genome of interest has been selected, results are summarized in tables. Each CRISPR is highlighted and *CAS* genes found in its vicinity are displayed. These are identified by dedicated HMM profiles we have built from available genes. If new putative *CAS* genes were found, there are highlighted in red. Annotations contains various elements such as positions, sequence of the consensus unit, links to external NCBI information, links to graphical circular view of genome (thanks to CGView, see (Stothard*et al.*, 2005)). Clicking on consensus jumps to the corresponding CRISPR's details and gives information such as units and spacers' coordinates, units and spacers' sequences, Pygram image (Durand *et al.*, 2006) and consensus WebLogo image (Fig. 1). Spacers, Units, CRISPR, flanking sequences, *CAS* genes may be downloaded in fasta format.



List of Crispr for genome Methanopyrus kandleri AV19

Genome info

Methanopyrus kandleri AV19
RefSeq : NC_003551
Sequence size : 1694969
4 CRISPR
Kingdom : Archaea
19 putatives genes CAS

Genomic region arround CRISPR

	BEGIN	END	
□	1275808	1277502	MK1278 hypothe
□	1292092	1292835	MK1292 fused nucleotidyltransf domain-containi

Fig. 1. Typical graphical views in CRISPI

2.2.2 Blast a personal sequence against CRISPI Microbiologists or virologists can Blast private sequences to find out if a virus or plasmid sequence matches with one or more spacers in the database. The query sequence must be in fasta format (DNA or protein). One can either paste the query sequence into the input field or upload it from a file on the local machine (files with multiple sequences are allowed). Users have access to Blast parameters for fine-tuned comparisons. Moreover, it can be run against units instead of spacers for studies on the origin of such structures. The Blast results pages are cross-linked with the CRISPI database so that it is easy to return to the database by clicking on hyperlinks.

2.2.3 Find out CRISPR structures in personal sequence People may want to check there own microbial sequence for annotation purpose. The query sequence must be in fasta format (only DNA sequences are allowed). One can either paste the query sequence into the input field or upload it from a file on the local machine files with multiple sequences are allowed). Results are summarized in a table with the option of being notified by email when available. These user-submitted genomes remain in confidential web pages that can be accessed for 10 days before deletion.

3 CONCLUSION

CRISPI is a dedicated environment on CRISPR in prokaryotic genomes that offers for the first time an up-to-date view of existing CRISPR (71 archaea totalling 291 CRISPR, and 987 bacteria totalling 2,103 CRISPR) including a complete repertory of CRISPR-associated genes -*CAS* genes-. The current version contains 1,173 archeal CAS genes and 4,396 bacterial CAS genes. We have not tried as in (Grissa *et al.*, 2007) to keep very small structures (1 or 2 spacers) as it is not clear if they have any relevance or activity. In contrast, we have included a richer environment for practical works on CRISPR : access to extended queries via Blast parameters, multiple graphical views, etc.

The next planned step in this work will be the automatic update of the database as new genomes will become available.

ACKNOWLEDGEMENTS

The necessary environment for all computations has been provided by the bioinformatics platform Biogenouest (http://biogenouest.org)

Funding: This work is supported by a grant from the ANR *Agence Nationale de la Recherche* through its MODULOME project.

REFERENCES

Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A. Romero, Philippe Horvath. (2007) CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes, *Science*, **315**, DOI: 10.1126/science.1138140.

Rotem Sorek, Victor Kunin, Philip Hugenholtz (2008) CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea, *Nature Reviews Microbiology*, **6**181-186, DOI:10.1038/nrmicro1793

Charles Bland, Teresa L Ramsey, Fareedah Sabree, Micheal Lowe, Kyndall Brown, Nikos C Kyrpides, Philip Hugenholtz (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats, *BMC Bioinformatics*, **8**:209, DOI:10.1186/1471-2105-8-209

Robert C. Edgar (2007) PILER-CR: Fast and accurate identification of CRISPR repeats, *BMC Bioinformatics*, **8**:18, DOI: 10.1186/1471-2105-8-18

Ibtissem Grissa, Gilles Vergnaud, Christine Pourcel (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats, *Nucleic Acids Research*, DOI:10.1093/nar/gkm360

Ibtissem Grissa, Gilles Vergnaud, Christine Pourcel (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats, *BMC Bioinformatics*, **8**:112, DOI:10.1186/1471-2105-8-172

Victor Kunin, Rotem Sorek, Philip Hugenholtz (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats, *Genome biology*, **5**:R61, DOI:10.1186/gb-2007-8-4-r61

Patrick Durand, Frédéric Mahé, Anne-Sophie Valin, Jacques Nicolas (2006) Browsing repeats in genomes: Pygram and an application to non-coding region analysis, *BMC Bioinformatics*, **7**:477, DOI:10.1186/1471-2105-7-477

Paul Stothard and David Wishart (2005) Circular genome visualization and exploration using CGView, *Bioinformatics*, **21**:4, DOI:10.1093/bioinformatics/bti054