CRISPI: a CRISPR Interactive database
Christine Rousseau, Mathieu Gonnet, Marc Le Romancer, Jacques Nicolas

To cite this version:

HAL Id: inria-00438512
https://hal.inria.fr/inria-00438512
Submitted on 7 Feb 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
CRISPI : a CRISPR Interactive database
Christine Rousseau 1,∗, Mathieu Gonnet 2, Marc Le Romancer 2, Jacques Nicolas 1,∗
1IRISA-INRIA, Campus de Beaulieu, 35042 Rennes cedex, France
2UMR 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 archaea 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 adaptive 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 questionable; A leader sequence is often mentioned 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
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 flanking sequences, CAS genome (thanks to CGView, see (Stothard et al., 2005)) to external NCBI information, links to graphical circular view of elements such as positions, sequence of the consensus unit, links found, there are highlighted in red. Annotations contains various CAS genes we have built from available genes. If new putative CAS genes were found, 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)).  

3 CONCLUSION

CRISPI is a dedicated environment on CRISPR in prokaryotic genomes that provides for the first time an up-to-date view of existing CRISPR (71 archaea totalling 291 CRISPR, and 987 bacteria totalling 2,105 CRISPR) including a complete repertory of CRISPR-associated genes -CAS genes-. The current version contains 1,173 archael CAS genes and 4,396 bacterial CAS genes.

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 Fremeaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A. Romero, Philippe Horvath. (2007) CRISPR - a widespread system that 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 viruses in prokaryotes and archaea, Nature Reviews Microbiology, 6:181-186, DOI: 10.1038/nrmicro1793


Bissem Grissa, Gilles Vergnaud, Christine POURCEL (2007) The CRISPRFinder 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

