CRISPRloci: Comprehensive and accurate annotation of CRISPR-Cas systems

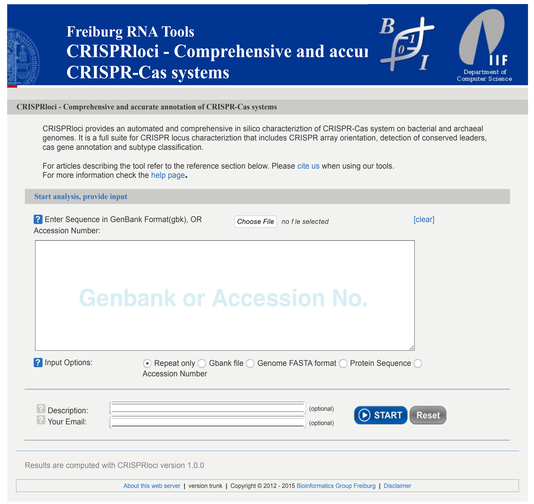
Omer S. Alkhnbashi[1], Shiraz A. Shah[2], Fabrizio Costa[1], Martin Mann[1], Xu Peng[2], Roger A. Garrett[2], Rolf Backofen[1]

[1]Bioinformatics Group, University of Freiburg, Freiburg, Germany. [2]Archaea Centre, University of Copenhagen Copenhagen Denmark.

INTRODUCTION

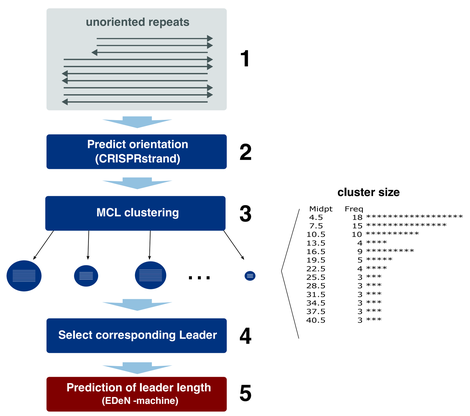
The CRISPR-Cas system is an adaptive immune system in archaea and bacteria, which provides resistance against invading viruses and plasmids. Identification of CRISPR-Cas systems on newly sequenced archaeal and bacterial genomes involves the correct definition and classification of both the coding and non-coding elements and this has always been challenging because of the high diversity and modularity of the systems. Thus, existing automated tools give only a partial definition of genomic CRISPR-Cas systems, and users are left to identify the remaining elements manually. We have developed a web-server called **CRISPRloci** for automated and comprehensive in silico characterization of CRISPR-Cas systems on archaeal and bacterial genomes. CRISPRloci visualizes the results in an interactive genome map and includes the ability to zoom in and click for additional information. CRISPR arrays are also classified into sequence families, or structural motifs, using our previous web-server CRISPRmap[1,2].

http://rna.informatik.uni-freiburg.de/



1. Omer S. Alkhnbashi, Fabrizio Costa, Shiraz A. Shah, Roger A. Garrett, Sita J. Saunders, and Rolf Backofen, *CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci.* Bioinformatics, 2014, 30(17), 489-496.
2. Sita J. Lange, Omer S. Alkhnbashi, Dominic Rose, Sebastian Will and Rolf Backofen, *CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems.* NAR, 2013, 41(17), 8034-8044.

METHODS



CRISPRloci integrates a series of tools in a seamless web interface featuring: (i) accurate prediction of all CRISPR arrays in the correct orientation; (ii) definition of CRISPR leaders for each locus with prediction of leader length using a machine learning method; (iii) annotation of *cas* genes and their unambiguous classification with respect to the official subtype classification using an accurate k-nearest neighbour clustering technique.

Although characterising the leader has always been a challenge due to low sequence conservation, a proper characterisation of the repeat will give clues to identifying the leader. We determine the location of the leader by first establishing the orientation of the repeat. Then we determine the length of the leader by using the repeat to fish for similar leaders across different hosts.

RESULTS

The CRISPRloci results page is divided into three main sections.

Right: Overview of CRISPR-Cas systems in the genome. Provides a global overview of CRISPR-Cas systems present in the genome and visualizes the results in an interactive genome map and includes the ability to zoom in and click for additional information.

CRISPR ID	Strand	Start Position	End Position	Consensus Repeat	Repeat Length	# of Repeats	Subtype
1	plus	16310	19901	CTTTCCTTCTACTAATCCCGGCGATCGGGACTGAAAC	37	50	I-D
2	plus	68499	72778	GTTCAACACCCTCTTTTCCCCGTCAGGGGACTGAAAC	37	57	III-C
3	plus	90105	92958	GTCTCCACTCGTAGGAGAAATTAATTGATTGGAAAC	36	39	III-B

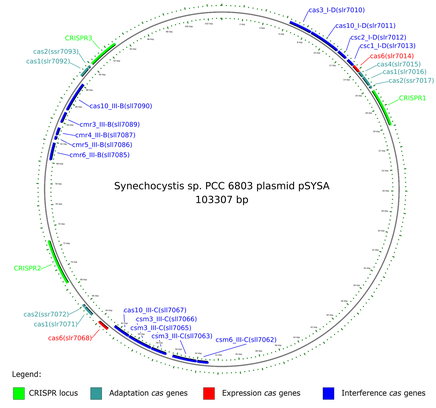
Feed the entire table data into CRISPRmap web server

Please select a row from above table to see more information about CRISPR system

CRISPR 1 at position 16310-19901 on plus strand with 37 repeat length and 50 number of repeats

Consensus: CTTTCCTTCTACTAATCCCGGCGATCGGGACTGAAAC

Feed this to CRISPRMap



Above: Table of CRISPR locus in the genome. Ordered list of CRISPR loci showing all the essential information, including strand and subtype. The list is clickable, revealing additional information about the locus of interest, including leader sequence, consensus repeat sequence and the option of forwarding this sequence to the CRISPRMap server, e.g. if a user wants to know which other organisms harbour similar CRISPRs.

Name	Annotation	Subtype	Function	Cassette	Strand	Start Position	End Position	Length
slr7011	cas10	I-D	Interference	1	plus	8531	11458	975
slr7012	csc2	I-D	Interference	1	plus	11524	12513	329
slr7013	csc1	I-D	Interference	1	plus	12674	13438	254
slr7014	cas6	I-D	Expression	1	plus	13441	14229	262
slr7015	cas4	I-D	Adaptation	1	plus	14222	14794	190
slr7016	cas1	I-D	Adaptation	1	plus	14804	15781	325
ssr7017	cas2	I-D	Adaptation	1	plus	15813	16097	94
sll7062	csm6	III-C	Interference	2	minus	52881	53996	371

Above: Table of cas genes annotation in the genome. Includes a proper annotation of *cas* genes, instead of a list of matching protein families from Pfam or CDD. Subtypes from the official classification are also listed along with the functional module that each gene belongs to. The sequence of the gene product can be obtained by clicking, which also reveals links to external databases like NCBI Gene, or Pfam.

CONCLUSION

CRISPRloci employs advanced machine learning techniques to accurately determine the Cas subtype, CRISPR orientation, leader location and extent, as well as proper annotation of *cas* genes, all of which have so far been missing from current online CRISPR resources. These features are presented in an interactive, clickable web interface which makes it easy for scientists to gain a full overview of the CRISPR systems in their organism of interest.