CRISPRloci: Comprehensive and accurate annotation of CRISPR-Cas systems

Omer S. Alkhnbashi\textsuperscript{1}, Shiraz A. Shah\textsuperscript{2}, Fabrizio Costa\textsuperscript{1}, Martin Mann\textsuperscript{1}, Xu Peng\textsuperscript{2}, Roger A. Garrett\textsuperscript{2}, Rolf Backofen\textsuperscript{1}

\textsuperscript{1}Bioinformatics Group, University of Freiburg, Freiburg, Germany. \textsuperscript{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 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\cite{1,2}.

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.

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.

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.