Towards an Automated Annotation of CRISPR-cas systems

Costa F[1], *Alkhnbashi OS*[1], *Shah SA*[2], *Saunders SJ*[1], *Barrangou R*[4], *Brouns SJJ*[5], *Charpentier E*[6], *Haft DH*[3], *Horvath P*[7], *Koonin EV*[3], *Makarova KS*[3], *Moineau S*[8], *Mojica FJM*[9], *Terns RM*[10], *Terns MA*[10], *van der Oost J*[5], *White MF*[11], *Wolf YI*[3], *Yakunin AF*[12], *Garrett RA*[2], *Backofen R*[1,13]

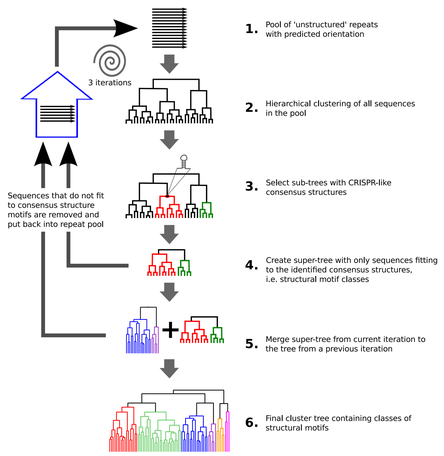
1. Bioinformatics group, Department of Computer Science, University of Freiberg, Georges-Kohler-Allee 106, 79110 Freiberg, Germany
2. Archaea Centre, Department of Biology, Copenhagen University, Ole Maaløes Vej 5, DK2200 Copenhagen N, Denmark
3. National Center for Biotechnology Information, NLM, National Institutes of Health, Bethesda, MD 20894, USA
4. Department of Food, Bioprocessing, and Nutrition Sciences, North Carolina State University, Raleigh, NC 27606, USA.
5. Laboratory of Microbiology, Wageningen University, Dreijenplein 10, 6703HB Wageningen,16 Netherlands
6. Department of Regulation in Infection Biology, Helmholtz Centre for Infection Research, D-38124 Braunschweig, Germany
7. DuPont Nutrition and Health, BP10, Dangé-Saint-Romain 86220, France.
8. Département de biochimie, de microbiologie et de bio-informatique, Faculté des sciences et de génie, Groupe de recherche en écologie buccale, Félix d'Hérelle Reference Center for Bacterial Viruses, Faculté de médecine dentaire, Université Laval, Québec City, Québec, Canada
9. Departamento de Fisiología, Genética y Microbiología. Universidad de Alicante. 03080-Alicante Spain
10. Biochemistry and Molecular Biology, Genetics and Microbiology, University of Georgia, Davison Life Sciences Complex, Green Street, Athens, GA 30602
11. Biomedical Sciences Research Complex, University of St Andrews, North Haugh, St Andrews, KY16 9TZ, UK
12. Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, M5S 3E5, Canada
13.BIOSS Centre for Biological Signaling Studies, Cluster of Excellence, University of Freiburg, Germany.

Automated annotationof CRISPR-cas loci in newly sequenced genomes is valuable both for general purposes of comparative genomics of archaea and bacteria, and for the progress of CRISPR research.
For a comprehensive analysis, several tasks have to be performed, among which the most important are:
1) the correct prediction of repeat orientation [CRISPRstrand]
2) a characterization of the repeats in terms of sequence and structure to infer repeat evolution [CRISPRmap]
3) annotation of the associated loci subtype according to the composition of cas genes.

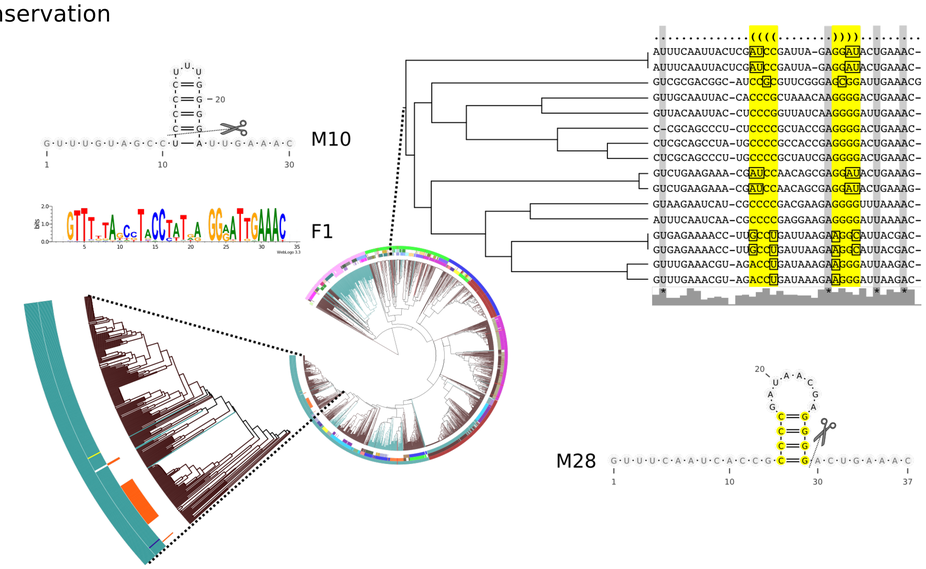
For the annotation of loci subtypes we present a solution that relies on a *novel similarity notion* for the interference modules. Loci annotation is achieved by nearest neighbor classification, which yields highly consistent results with respect to the current subtype classification.

CRISPRmap: an automated classification of repeat conservation

Basic procedure for identifying conserved hairpin motifs in all repeats.



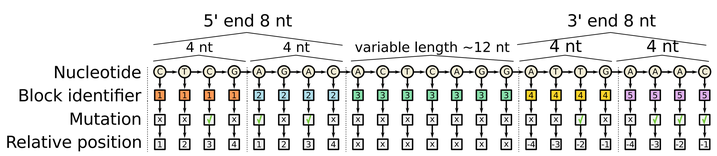
We predicted CRISPR loci in all publically available genome sequences using the CRT tool and CRISPRfinder.
In total, we analysed 4500 CRISPR loci in 4899 genomes:
4590 bacteria, 309 archaea



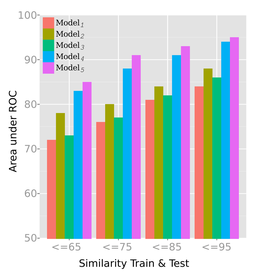
Highlighting the advantage of independent clustering approaches. (A) CRISPRs in the largest sequence family, F1, are mostly unstructured; however, for 50 CRISPRs also a conserved structure motif, M10, was identified.
This indicates that subsets of conserved families can bestructured. F1 contains the conserved 50 tag, marked with the magenta box. (B) Structure motif M28 shows no sequence conservation, but a conserved structure (base pairs are highlighted in yellow). The many compensatory base pairs are marked in the alignment with squares.
This structure has been verified via mutational analyses. Potential cleavage sites are indicated as observed in the literature

CRISPRstrand: predicting repeat orientations at CRISPR loci

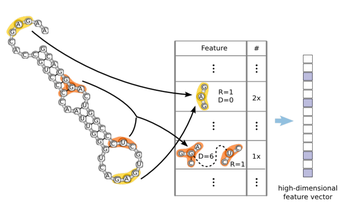
Although existing bioinformatics tools can recognize CRISPR loci by their characteristic repeat-spacer architecture, they generally output CRISPR arrays of ambiguous orientation and thus do not determine the strand from which crRNAs are processed.
Knowledge of the correct orientation is crucial for many tasks, including the classification of CRISPR conservation, the detection of leader regions, the identification of target sites (protospacers) on invading genetic elements and the characterization of protospacer-adjacent motifs.



Graph encoding the consensus repeat sequence. The consensus nucleotide information is represented as a path graph, and additional informationis modelled as a chain of additional vertices: block identifiers, mutation and relative position in the block.

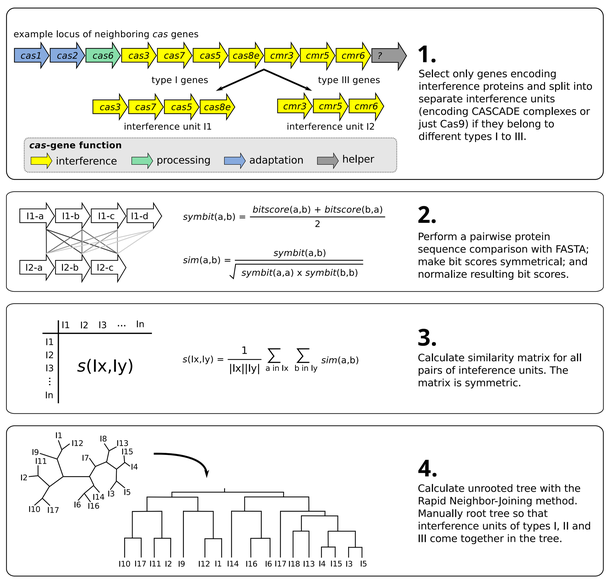


AUC ROC performance comparison of the five models thatencode increasing amount of information about the CRISPR arrays.



Generic machine learning approach for structured data: decompose object into fragments (like k-mers for sequences) and use their occurrence counts as features.

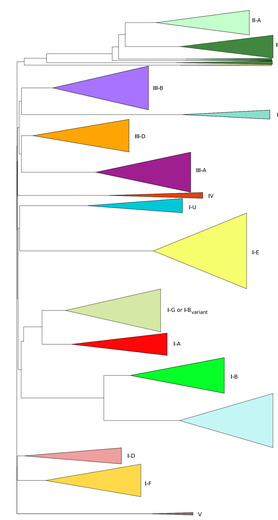
Clustering of CRISPR-Cas crRNA-effector modules by protein sequence similarity



Given the variability of the gene composition and architecture of the effector modules and the lack of a universal marker suitable for phylogenetic analysis, we developed a simple clustering approach based on sequence similarity between the proteins of these modules.

Cas loci can contain multiple modules of different types, we therefore separated them accordingly. Gene-to-gene similarities were calculated and aggregated into a locus-wide value. The resulting cluster dendrogram showed a high correlation with the subtype classification.

Using prior information on the association between sequence profiles and CRISPR-cas loci and the corresponding classification of the effector modules, we developed a k-nearest neighbor classifier that achieved 0.998 accuracy.



Dendrogram of Cas loci using proposed similarity notion