Towards an Automated Annotation of CRISPR-cas systems

Costa F1, Alkhnabshi OS2, Shah SA3, Saunders SJ4, Barrangou R4, Brouns SJ5, Charpentier E6, Haft DH7, Horvath P8, Koonin EV9, Makarova KS10, Moineau S11, Mojica FJM12, Terns RM13, Terns MA13, van der Oost J14, White MF15, Wolf YI16, Yakunin AF17, Garrett RA12, Backofen R12,13

Automated annotation of 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.

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) as well as the design of protospacer-adjacent motifs.

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 those 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-wise 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.