Mechanisms of RNA-mediated regulation

CRISPR revisited: structure prediction of CRISPR repeats

Sita Lange[1], Omer S. Alkhnbashi [1], Dominic Rose[1], Ingeborg Scholz[2], Wolfgang Hess[2], Rolf Backofen[1]

[1]Bioinformatics Group, University of Freiburg

[2]Genetics & Experimental Bioinformatics, University of Freiburg

Biological function

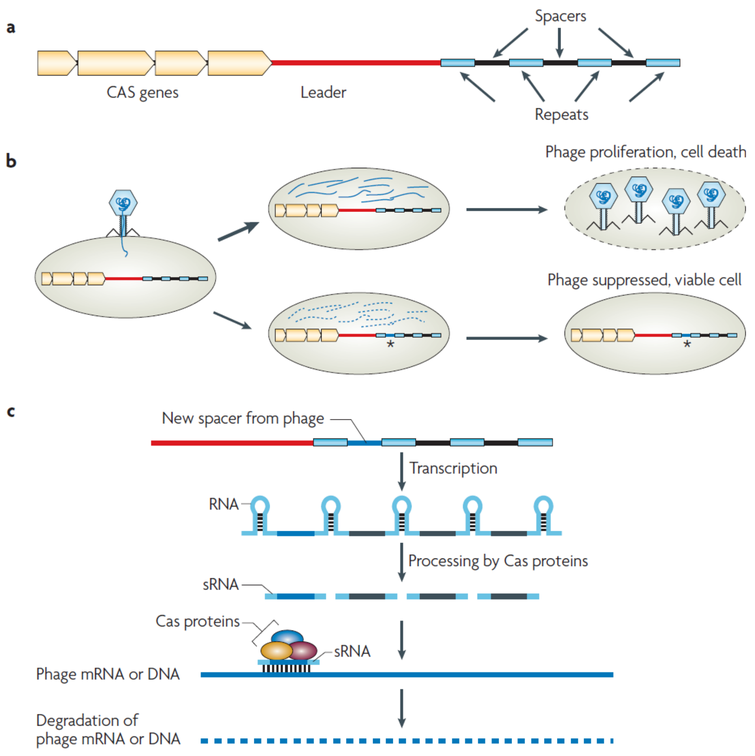
Prokaryotes are known to acquire immunity against phages and viruses through a widely conserved RNA-based gene silencing pathway. This mechanism involves the non-coding RNA, called Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs), illustrated to the right.

The CRISPR transcripts consist of a leader sequence followed by an array of alternating repeat and spacer sequences. The spacer sequences have been found to match foreign virus or plasmid DNA.

A set of CRISPR-associated (Cas) proteins is often located in close vicinity of the CRISPR array. These proteins guide the targeting mechanism and acquire additional spacer sequences from new invaders.

One Cas protein (endonuclease) binds to the CRISPR array at a stem-loop motif in the repeats and cleaves the precursor into small mature crRNAs, ready for targeting.

CRISPR mechanism



Sorek *et al.*, Nature Reviews Microbiology 2008

Summary

We apply two approaches to characterise the sequence-structure motif to which the Cas protein binds.

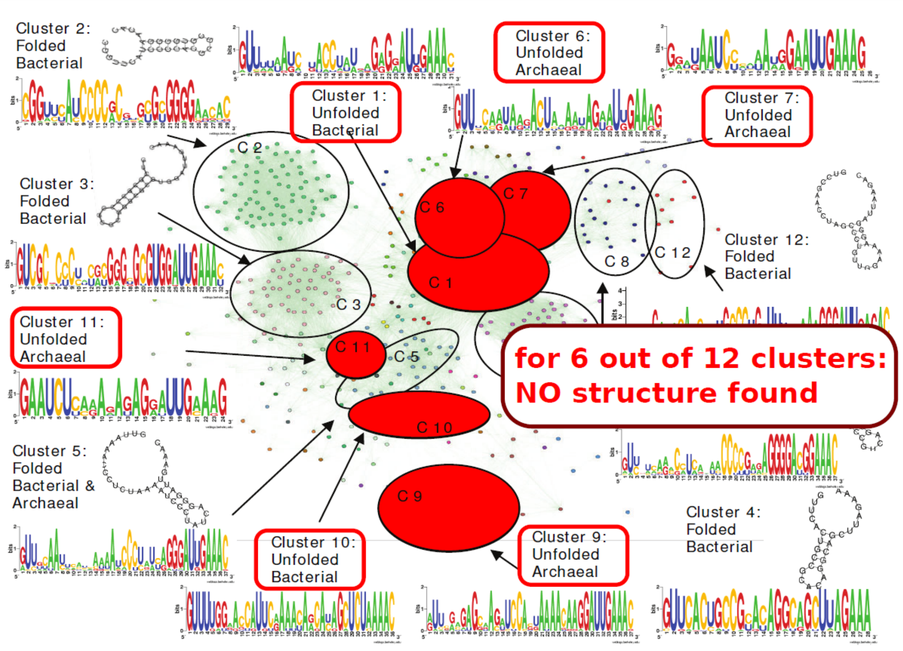
(1) Large-scale clustering of all CRISPR repeats into sequence-structure families.

(2) Analysis of the repeat structure in the context of the spacer sequences.

We compare our results from structure-based clustering to previously published sequence-based clusters. We have already identified 4 novel structured classes and expect to detect more.

A substantial effect on the repeat structure is visible when considering the context during structure prediction. We highlight that the minimal free energy structure is not always the most dominant. We further propose a hypothetical processing order by investigating the accuracy of structure candidates at each repeat position.

Classes of sequence-structure families



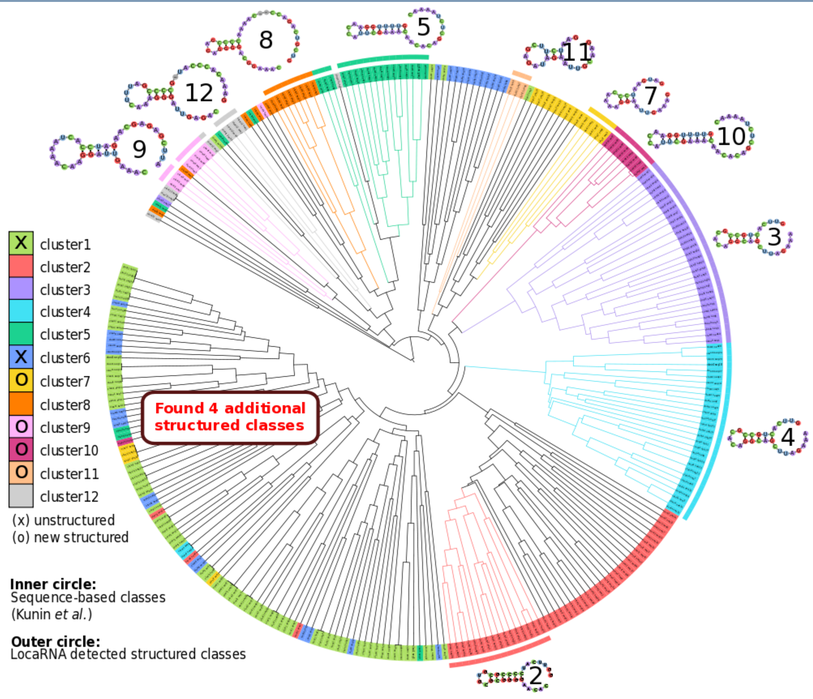
Kunin *et al.*, Genome Biology 2007

Left: A modified figure from Kunin *et al*. to highlight the 6 out of 12 classes for which no common structure have been reported.

Right: We re-clustered the same CRISPR repeats with LocaRNA and find structures for 4 of the previously 6 unstructured classes.

In many cases, the consensus structure only includes a subset of the CRISPR repeats found in the original classes.

In future clustering approaches we will consider the orientation and include the influence of the precursor context on the repeat structure (see below).



LocaRNA: Will *et al.*, PLOS Computational Biology 2007

CRISPR structure prediction: the influence of context sequences

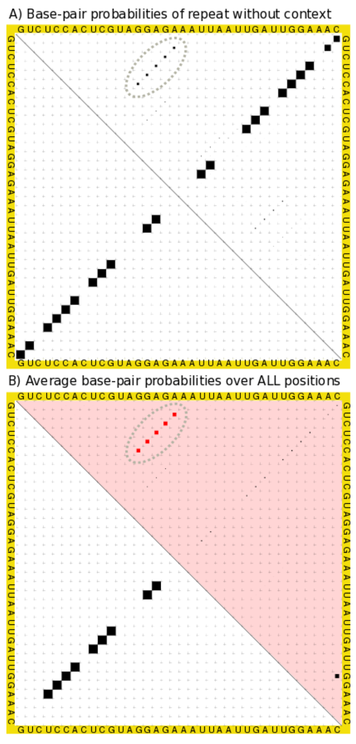
Right: The dotplots efficiently illustrate the influence of the context on the repeat structure. Each dot in the lower triangle represents a base-pair in the minimum free energy (MFE) structure. The dots in the upper triangle present:

A) Base-pair probabilities of the structure ensemble of the direct repeat folded by RNAfold.

B) Average base-pair probabilities over all repeat positions in the entire CRISPR array, folded locally with RNAplfold.

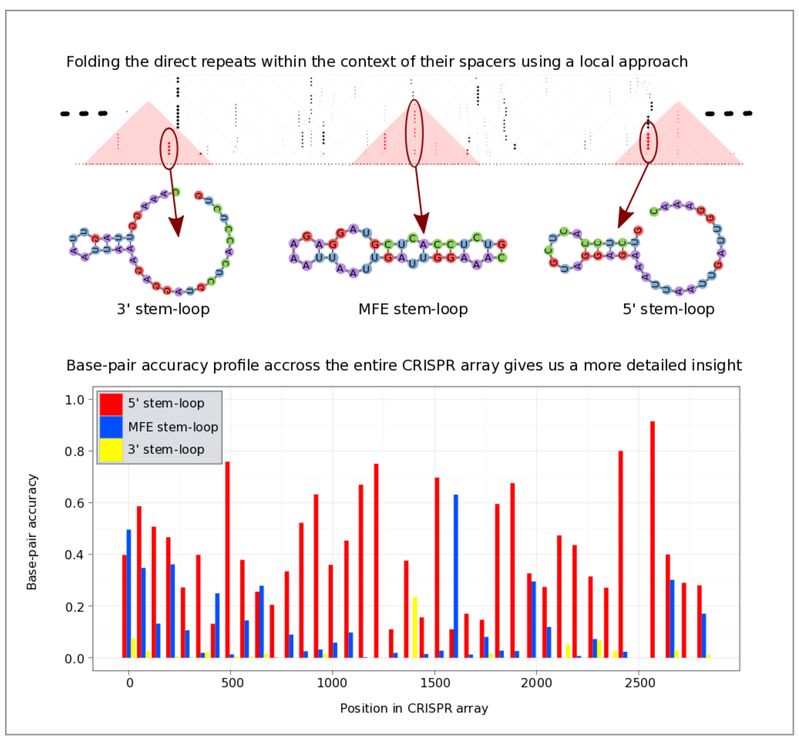
RNAfold: Hofacker *et al.*, Monatshefte für Chemie 1994

RNAplfold: Bernhart *et al.*, Bioinformatics 2006



Right: We see a section of the dotplot from the local folding with RNAplfold and 3 candidate CRISPR structures. To identify the most likely structure motif we analyze the accuracy profile of the CRISPR array below. Each bar presents the mean base-pair probability of a candidate. This profile may help in determining a processing order of the mature crRNA, i.e. the order of RNA cleavages.

We see clearly that the MFE structure nearly disappears when folded within the context. Instead, the 5bp stem-loop at the 5'-end of the repeat becomes more dominant in the context of the spacers. In addition, its shape fits best to previously published CRISPR stem-loop structures.



gefördert durch die:

DFG

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

FORSYS Research Units for Systems Biology

bioss

zbsa zentrum für biosystemanalyse